Helsinki University of Technology Laboratory of Computational Engineering Publications
Teknillisen korkeakoulun Laskennallisen tekniikan laboratorion julkaisuja
Espoo 2006 REPORT B54

RECURSIVE BAYESIAN INFERENCE ON STOCHASTIC DIFFERENTIAL EQUATIONS

Simo Särkkä

$d\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\, dt + \mathbf{L}(t)\, d\boldsymbol{\beta}$
$p(\mathbf{x} \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid \mathbf{x})\, p(\mathbf{x})}{p(\mathbf{y})}$

TEKNILLINEN KORKEAKOULU
TEKNISKA HÖGSKOLAN
HELSINKI UNIVERSITY OF TECHNOLOGY
TECHNISCHE UNIVERSITÄT HELSINKI
UNIVERSITE DE TECHNOLOGIE D'HELSINKI

Helsinki University of Technology Laboratory of Computational Engineering Publications
Teknillisen korkeakoulun Laskennallisen tekniikan laboratorion julkaisuja
Espoo 2006 REPORT B54

RECURSIVE BAYESIAN INFERENCE ON STOCHASTIC DIFFERENTIAL EQUATIONS

Simo Särkkä

Dissertation for the degree of Doctor of Science in Technology to be presented with due permission of the Department of Electrical and Communications Engineering, Helsinki University of Technology, for public examination and debate in Auditorium S4 at Helsinki University of Technology (Espoo, Finland) on the 24th of April, 2006, at 12 noon.

Helsinki University of Technology
Department of Electrical and Communications Engineering
Laboratory of Computational Engineering

Teknillinen korkeakoulu
Sähkö- ja tietoliikennetekniikan osasto
Laskennallisen tekniikan laboratorio

Distribution:
Helsinki University of Technology
Laboratory of Computational Engineering
P. O. Box 9203
FIN-02015 HUT
FINLAND
Tel. +358-9-451 4826
Fax. +358-9-451 4830
http://www.lce.hut.fi

Online in PDF format: http://lib.hut.fi/Diss/2006/isbn9512281279/

E-mail: Simo.Sarkka@hut.fi

ISBN 951-22-8126-0 (printed)
ISBN 951-22-8127-9 (PDF)
ISSN 1455-0474
PicaSet Oy
Espoo 2006

Abstract

This thesis is concerned with recursive Bayesian estimation of non-linear dynamical systems, which can be modeled as discretely observed stochastic differential equations. The recursive real-time estimation algorithms for these continuous-discrete filtering problems are traditionally called optimal filters and the algorithms for recursively computing the estimates based on batches of observations are called optimal smoothers. In this thesis, new practical algorithms for approximate and asymptotically optimal continuous-discrete filtering and smoothing are presented.

The mathematical approach of this thesis is probabilistic and the estimation algorithms are formulated in terms of Bayesian inference. This means that the unknown parameters, the unknown functions and the physical noise processes are treated as random processes in the same joint probability space. The Bayesian approach provides a consistent way of computing the optimal filtering and smoothing estimates, which are optimal given the model assumptions and a consistent way of analyzing their uncertainties.

The formal equations of the optimal Bayesian continuous-discrete filtering and smoothing solutions are well known, but the exact analytical solutions are available only for linear Gaussian models and for a few other restricted special cases. The main contributions of this thesis are to show how the recently developed discrete-time unscented Kalman filter, particle filter, and the corresponding smoothers can be applied in the continuous-discrete setting. The equations for the continuous-time unscented Kalman-Bucy filter are also derived.

The estimation performance of the new filters and smoothers is tested using simulated data. Continuous-discrete filtering based solutions are also presented to the problems of tracking an unknown number of targets, estimating the spread of an infectious disease and to prediction of an unknown time series.

Tiivistelmä

Tämä väitöskirja käsittelee diskreetein väliajoin havaittavina stokastisina differentiaaliyhtälöinä mallinnettavissa olevien epälineaaristen dynaamisten järjestelmien rekursiivista bayesilaista estimointia. Tällaisten jatkuva-diskreettien suodatusmallien rekursiivisia reaaliaikaisia estimointialgoritmeja kutsutaan perinteisesti optimaalisiksi suotimiksi ja algoritmeja, jotka laskevat rekursiivisesti estimaatteja käyttäen kokonaisia datajoukkoja kutsutaan optimaalisiksi silottimiksi. Tässä väitöskirjassa esitetään uusia käytännöllisiä algoritmeja likimääräiseen ja asymptoottisesti optimaaliseen jatkuva-diskreettiin suodatukseen sekä siloitukseen.

Väitöskirjan matemaattinen lähestymistapa on probabilistinen ja estimointialgoritmit formuloidaan bayesilaista päättelyä käyttäen. Tämä tarkoittaa sitä, että tuntemattomat parametrit, tuntemattomat funktiot sekä fysikaaliset kohinaprosessit käsitellään satunnaisprosesseina samassa yhteisessä todennäköisyysavaruudessa. Bayesilainen lähestymistapa tarjoaa yhtenäisen tavan laskea optimaalisia suodatus- ja silotus-estimaatteja, jotka ovat optimaalisia mallioletusten valossa, sekä yhtenäisen tavan niiden epävarmuuksien analysointiin.

Optimaalisen bayesilaisen jatkuva-diskreetin suodatuksen ja silotuksen formaalit yhtälöt ovat laajasti tunnettuja, mutta tarkat analyyttiset ratkaisut ovat olemassa vain lineaaris-gaussisille malleille sekä muutamalle muulle rajoitetulle erikoistapaukselle. Tämän väitöskirjan pääkontribuutiona on osoittaa, kuinka vähän aikaa sitten kehitettyjä diskreettiaikaisia hajustamattomia Kalmanin suotimia, partikkelisuotimia sekä vastaavia silottimia voidaan soveltaa jatkuva-diskreetissä tapauksessa. Myös jatkuva-aikaisen hajustamattoman Kalmanin-Bucyn suotimen yhtälöt johdetaan.

Uusien suotimien ja silottimien suorituskyky testataan simuloidulla datalla. Jatkuva-diskreetteihin suotimiin perustuvat ratkaisut esitetään myös kohteiden seurantaan tapauksessa, jossa kohdemäärä on tuntematon, tarttuvan taudin leviämisen estimointiin sekä tuntemattoman aikasarjan ennustamiseen.

Preface

This thesis is the result of my research work in the Laboratory of Computational Engineering at Helsinki University of Technology during the years 2001 to 2006. The work was funded by the Academy of Finland Centre of Excellence in Computational Science and Engineering, and the SmartSpace1 project, coordinated by the Technical Research Centre of Finland VTT and funded by the National Technology Agency TEKES.

I with to express my sincere gratitude to Prof. Jouko Lampinen for his excellent instruction and supervision of my thesis, and to Dr. Aki Vehtari for encouraging support and advice in the research, especially in everything related to the Bayesian inference. I also wish to thank Prof. Kimmo Kaski for providing first rate computational resources and the premises for the research work. I also wish to thank Dr. Tommi Sottinen and Dr. Marko Vauhkonen for reviewing the thesis and for their comments and suggestions on the text.

I am grateful to Dr. Toni Tamminen, Lic. Tech. Matti Vihola, M. Sc. Aapo Nummenmaa, M. Sc. Ilkka Kalliomäki and Dr. Jari Saramäki for research collaboration and their help with various parts of the thesis. For providing insight into practical applications of optimal filtering and for helping in defining the objectives of the research I wish to thank Dr. Petri Korpisaari, Dr. Petri Salmela, Lic. Tech. Kari Heine, Dr. Amaury Lendasse, M. Sc. Hannu Maula and M. Sc. Juhani Korkealaakso.

The people at Indagon Ltd., especially Mr. Mikko Weckström and Mr. Markus Mikkolainen deserve thanks for providing insight into the practical issues related to optimal filtering. For helping with work facilities and practicalities I wish to thank Ms. Eeva Lampinen and Ms. Aino Järvenpää. I would also like to thank all the people at the Laboratory of Computational Engineering for the inspiring research environment.

Finally, I wish to thank my lovely wife Susanne for supporting me during these years. I would also like to thank all the friends and family for helping me to maintain a life outside the work.

Simo Särkkä

Notation

General Notation

$a, b, c, x, t, \alpha, \beta$	Scalars
$\mathbf{a}, \mathbf{f}, \mathbf{s}, \mathbf{x}, \mathbf{y}, \boldsymbol{\alpha}, \boldsymbol{\beta}$	Vectors
$\mathbf{A}, \mathbf{F}, \mathbf{S}, \mathbf{X}, \mathbf{Y}$	Matrices
$\mathcal{A}, \mathcal{F}, \mathcal{S}, \mathcal{X}, \mathcal{Y}$	Sets
$\mathscr{A}, \mathscr{F}, \mathscr{S}, \mathscr{X}, \mathscr{Y}$	Algebras
$\mathbb{A}, \mathbb{F}, \mathbb{S}, \mathbb{X}, \mathbb{Y}$	Spaces
$\mathpzc{A}, \mathpzc{F}, \mathpzc{S}, \mathpzc{X}, \mathpzc{Y}$	Operators and functionals

Notational Conventions

$\frac{\mathrm{d}\mathbf{g}}{\mathrm{d}t}$	Derivative of function $\mathbf{g}(t)$ with respect to t
$\frac{\partial g_i(\mathbf{x})}{\partial x_j}$	Partial derivative of g_i with respect to x_j
$\operatorname{diag} \mathbf{A}$	Diagonal of matrix A
$\operatorname{diag}(a_1, \ldots, a_n)$	Diagonal matrix with diagonal values $a_1, \ldots, a_n$
$\mathrm{E}[\mathbf{x}]$	Expectation of $\mathbf{x}$
$\mathrm{E}[\mathbf{x} \mid \mathbf{y}]$	Expectation of $\mathbf{x}$ given $\mathbf{y}$
$p(\mathbf{x})$	Probability density of continuous random variable $\mathbf{x}$ or point probability of discrete random variable $\mathbf{x}$
$p(\mathbf{x} \mid \mathbf{y})$	Conditional probability (density or point) of $\mathbf{x}$ given $\mathbf{y}$
$P(\mathbf{x})$	Probability of discrete random variable $\mathbf{x}$
$P(\mathbf{x} \mid \mathbf{y})$	Conditional probability of discrete random variable $\mathbf{x}$
$p(\mathbf{x}) \propto q(\mathbf{x})$	$p(\mathbf{x})$ is proportional to $q(\mathbf{x})$
$\operatorname{tr} \mathbf{A}$	Trace of matrix A
$\mathbf{x} \sim p(\mathbf{x})$	Random variable $\mathbf{x}$ has the distribution $p(\mathbf{x})$
$x \triangleq y$	x is defined to be equal to y
$x \gg y$	x is much greater than y
$\dot{\mathbf{x}}$	Time derivative of $\mathbf{x}(t)$
$x_{j,k}^{(i)}$	The element j of the vector $\mathbf{x}$, which is related to the particle i on time step k

Symbols

α	Parameter of unscented transform
a	Angular velocity or unknown parameter
$\mathbf{a}$	Action or decision in statistical decision theory
$\mathbf{a}_o$	Optimal action or decision
$\mathbf{a}(t)$	Action or decision function
$\mathbf{a}_o(t)$	Optimal action or decision function
$\mathbf{A}$	Lower triangular Cholesky factor of covariance matrix
$\mathbf{A}(t)$	Transition matrix
$\mathbf{A}_k$	Dynamic model matrix (i.e., transition matrix) of jump from step k to step $k+1$
$\mathscr{A}$	σ-algebra on the space of events Ω
$\mathcal{A}_t[\cdot]$	Characteristic operator of Itô diffusion
$\mathcal{A}_t^*[\cdot]$	Kolmogorov forward (i.e., Fokker-Planck) operator
β	Parameter of unscented transform
$\beta(t)$	Scalar Brownian motion
$\boldsymbol{\beta}(t)$	Multidimensional Brownian motion
$\mathbf{B}$	Dispersion matrix of importance process
$\mathscr{B}[0,\infty)$	Borel σ-algebra on $[0,\infty)$
c	Scale parameter of matrix form unscented transform
c_k	Data association indicator on time step k
$\mathbf{C}(t)$	Time varying matrix in matrix fraction decomposition
$\mathbf{C}_k$	Gain matrix in optimal smoother or cross-covariance matrix in optimal filter/smoother
$\mathbf{C}_L$	Cross-covariance in linear approximation of a non-linear transform
$\mathbf{C}_Q$	Cross-covariance in quadratic approximation of a non-linear transform
$\mathbf{C}_U$	Cross-covariance in unscented approximation of a non-linear transform
$\delta(\cdot)$	Dirac delta function
$\delta_i(\cdot)$	Detection indicator
$\partial^2(\cdot)$	Second order correction term in quadratic approximation of SDE
$\mathrm{d}\boldsymbol{\beta}(t)$	Differential of Brownian motion
$\delta\mathbf{x}$	Small but finite perturbation of vector $\mathbf{x}$
δt	Small but finite perturbation of time variable t
Δt	Small but finite perturbation of time variable t
$\Delta\boldsymbol{\beta}$	Increment of Brownian motion on finite time interval
$\mathrm{d}t$	Differential of time variable t
$\mathrm{d}\mathbf{x}$	Differential (stochastic or deterministic) of vector $\mathbf{x}$
$\mathbf{D}(t)$	Time varying matrix in matrix fraction decomposition
$\mathbf{D}_k$	Gain matrix in optimal smoother

$\mathbf{e}(t)$	Gaussian white noise process, which is defined as the formal derivative of Brownian motion $\mathbf{e}(t) = \mathrm{d}\boldsymbol{\eta}(t)/\mathrm{d}t$
e_k	Gaussian random variable
$\mathbf{e}_i$	Unit vector in the direction of the coordinate axis i
$\mathbf{e}_k$	Visibility indicator on time step k
$f(t, \omega)$	Stochastic process
$\mathbf{f}(\cdot)$	Drift function of stochastic differential equation or transition function in discrete-time dynamic model
$\mathbf{F}(t)$	Feedback matrix of linear stochastic differential equation
$\mathbf{F}_{\mathbf{x}}(\cdot)$	Jacobian matrix of function $\mathbf{f}(\cdot)$
$\mathbf{F}^{(i)}_{\mathbf{x}\mathbf{x}}(\cdot)$	Hessian matrix of $f_i(\cdot)$
$\mathscr{F}_t$	Natural filtration of Brownian motion
$\phi_n(\cdot)$	Simple stochastic process
$\mathbf{g}(\cdot)$	Some anonymous function or drift function of importance process
$\mathbf{G}_{\mathbf{x}}(\cdot)$	Jacobian matrix of function $\mathbf{g}(\cdot)$
$\mathbf{G}^{(i)}_{\mathbf{x}\mathbf{x}}(\cdot)$	Hessian matrix of $g_i(\cdot)$
$\mathscr{H}_t$	Anonymous filtration
$\mathbf{H}_k$	Measurement matrix on time step k in linear Gaussian model
$\mathbf{H}_{\mathbf{x}}(\cdot)$	Jacobian matrix of function $\mathbf{h}(\cdot)$
$\mathbf{H}^{(i)}_{\mathbf{x}\mathbf{x}}(\cdot)$	Hessian matrix of $h_i(\cdot)$
$\mathcal{H}$	Set of hypotheses
$\mathcal{H}(\cdot)$	Anonymous functional
$\mathbf{h}(\cdot)$	Measurement model function or an arbitrary function
i	Integer valued index variable
$\mathbf{I}$	Identity matrix
$\mathcal{I}[\cdot]$	Itô integral of argument
$\mathbf{J}(\cdot)$	Jacobian matrix
κ	Parameter of unscented transform
k	Discrete time step number
$\mathbf{K}_k$	Kalman gain matrix in optimal filter
$\mathbf{K}_c(t)$	Kalman gain matrix in continuous-time optimal filter
λ	Parameter of unscented transform
$\lambda(t)$	Logarithm of likelihood ratio
$\boldsymbol{\lambda}_k$	Latent variable
$\mathbf{L}(\cdot)$	Dispersion matrix of stochastic differential equation
$L(\cdot)$	Loss function or likelihood function
m	Positive integer, usually the dimensionality of measurement
$\mathbf{m}$	Mean of Gaussian distribution
m_k	Mean of scalar discrete-time Gaussian process on time step k
$\mathbf{m}_k$	Mean of discrete-time Gaussian process on time step k
$\mathbf{m}_k^-$	Predicted mean just before measurement $\mathbf{y}_k$
$\mathbf{m}_k^s$	Mean computed by optimal smoother

$\mathbf{m}(t)$	Mean of a Gaussian process as function of time
$\mathbf{m}^s(t)$	Mean computed by optimal smoother
$\boldsymbol{\mu}_k$	Predicted mean of measurement $\mathbf{y}_k$
$\boldsymbol{\mu}_L$	Mean in linear approximation of a non-linear transform
$\boldsymbol{\mu}_Q$	Mean in quadratic approximation of a non-linear transform
$\boldsymbol{\mu}_U$	Mean in unscented approximation of a non-linear transform
$\boldsymbol{\eta}(t)$	Brownian motion (in continuous-time measurement model)
n	Positive integer, usually the dimensionality of state
$\mathbf{n}(t)$	Brownian motion
N	Positive integer, usually number of Monte Carlo samples
$\mathrm{N}(\cdot)$	Normal distribution
$o(g(\Delta t))$	Function such that $o(g(\Delta t))/g(\Delta t) \to 0$ when $\Delta t \to 0$
$O(g(\Delta t))$	Function such that $O(g(\Delta t))/g(\Delta t) \to$ constant when $\Delta t \to 0$
$\pi(\cdot)$	Importance distribution
p_k	Variance of scalar discrete-time Gaussian process on time step k
P	Probability measure
$\tilde{P}$	Transformed probability measure
$\mathbf{P}$	Covariance of Gaussian distribution
$\mathbf{P}_k$	Covariance of discrete-time Gaussian process on time step k
$\mathbf{P}_k^-$	Predicted covariance just before measurement $\mathbf{y}_k$
$\mathbf{P}_k^s$	Covariance computed by optimal smoother
$\mathbf{P}(t)$	Covariance of a Gaussian process as function of time
$\mathbf{P}^s(t)$	Covariance computed by optimal smoother
q	Spectral density of scalar white noise process, diffusion coefficient of scalar Brownian motion or variance of scalar Gaussian process noise
$q(t)$	Time varying spectral density of scalar white noise process or diffusion coefficient of scalar Brownian motion
$\mathbf{q}_k$	Discrete-time Gaussian process noise
$\mathbf{Q}(t)$	Discretized process noise covariance
$\mathbf{Q}_k$	Covariance of discrete-time process noise at jump from step k to $k+1$
$\mathbf{Q}_c(t)$	Diffusion matrix of a Brownian motion or spectral density of the corresponding white noise process
$\mathbf{R}_k$	Covariance matrix of measurement at step k
$\mathbf{R}_c(t)$	Diffusion matrix of a Brownian motion or spectral density of a white noise process (in continuous-time measurement model)
$\mathbb{R}$	Space of real numbers
$\mathbb{R}_+$	Space of positive real numbers $[0, \infty)$
$\mathbb{R}^n$	n-dimensional space of real numbers
$\mathbb{R}^{n \times m}$	Space of real $n \times m$ matrices
r	Variance of measurement noise

r_k	Range measurement at time step k
$\mathbf{r}_k$	Discrete-time Gaussian measurement noise
σ^2	Variance
$\sigma(\cdot)$	The minimal σ-algebra
s	Time variable $s \in [0, \infty)$ or the dimensionality of process noise
$\mathbf{s}(t)$	Importance process
$\mathbf{s}^*(t)$	Normalized importance process
$\mathcal{S}[\cdot]$	Stratonovich integral of argument
S	Index of the first time step or the start time of a time interval
$\mathbf{S}_k$	Innovation covariance in optimal filter
$\mathbf{S}_L$	Covariance in linear approximation of a non-linear transform
$\mathbf{S}_Q$	Covariance in quadratic approximation of a non-linear transform
$\mathbf{S}_U$	Covariance in unscented approximation of a non-linear transform
θ_k	Angular measurement on time step k
$\theta_j(\omega)$	A scalar random variable
$\boldsymbol{\theta}$	A random variable
$\boldsymbol{\theta}(t)$	A random process
τ	Time variable $\tau \in [0, \infty)$
t	Time variable $t \in [0, \infty)$
t_k	Time instance of time step k (usually of measurement y_k)
T	Index of the last time step, the final time of a time interval or the number of targets in multiple target tracking
T_k	Number of targets on time step k
$\mathbf{T}_k$	Sufficient statistics
$\mathbf{u}(t)$	Known deterministic function of time (input function)
$U(\cdot)$	Utility function
$\mathbf{v}_k$	Innovation vector in optimal filter
V	Volume of measurement space
$\mathbf{V}(\cdot)$	Dispersion matrix of stochastic differential equation in continuous-time measurement model
ω	Outcome (event) of random experiment
Ω	Space of events ω
$w^{(i)}$	Weight of particle i in importance sampling
$\mathbf{w}(t)$	Gaussian white noise process, which is defined as the formal derivative of Brownian motion $\mathbf{w}(t) = \mathrm{d}\boldsymbol{\beta}(t)/\mathrm{d}t$
$W_i^{(m)}$	Mean weight of unscented transform
$W_i^{(c)}$	Covariance weight of unscented transform
$\mathbf{w}_m$	Mean weight vector in matrix form unscented transform
$\mathbf{W}$	Covariance weight matrix in matrix form unscented transform
$\mathbf{x}$	Random variable
$\mathbf{x}_k$	State on time step k
$\mathbf{x}_{1:k}$	Set containing the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_k$

$\mathbf{x}(t)$	State at time instance t
$\mathbf{X}$	Matrix of sigma points of $\mathbf{x}$
$\mathscr{X}_t$	The natural filtration of stochastic process $\mathbf{x}(t;\omega)$
$\mathcal{X}(\omega)$	Stochastic process
$\mathbf{y}_k$	Measurement at time step k
$\mathbf{y}_{1:k}$	Set containing the vectors $\mathbf{y}_1,\ldots,\mathbf{y}_k$
$\mathbf{y}(t)$	Continuous-time measurement at time instance t
$\mathbf{Y}$	Matrix of sigma points of $\mathbf{y}$
$\mathscr{Y}_t$	The natural filtration of stochastic process $\mathbf{y}(t;\omega)$
$\mathbf{z}(t)$	Formal derivative of continuous-time measurement $\mathbf{z}(t) = \mathrm{d}\mathbf{y}(t)/\mathrm{d}t$
$Z(\cdot)$	Likelihood ratio of stochastic processes
Z_k	Normalization constant of filtering distribution
∞	Infinity

Contents

Chapter 1

Introduction

1.1 Purpose and Contributions of the Thesis

The purpose of this thesis is to provide new algorithms for optimal continuous-discrete filtering and smoothing, that is, new algorithms for recursive estimation of states of systems, which can be modeled as discretely observed stochastic differential equations. The emphasis is in stochastic differential equations with non-linear drift terms and measurement models with non-linear and non-Gaussian components, where the classical linear Kalman filter cannot be used. These kind of non-linear continuous-discrete filtering problems have previously been solved using Taylor series expansion based approximation methods (extended Kalman filters), but other types of methods have less been developed for the continuous-discrete case. In this thesis, it is shown how more recent discrete-time filtering algorithms, the unscented Kalman filter and particle filters can be modified for use in continuous-discrete filtering problems.

The mathematical treatment of the models and algorithms in this thesis is entirely Bayesian, which means that all the results are treated as being approximations to certain probability distributions or their parameters. Probability distributions are used for modeling both the uncertainties in the models and for modeling the physical randomness. In this thesis, the theory of probabilistic non-linear optimal filtering is formulated in terms of Bayesian inference and both the classical and recent filtering algorithms are reviewed using the same notation and formalism. The emphasis is in the continuous-discrete filtering problems, but also the pure discrete-time (discrete-time dynamics, discrete-time measurements) and pure continuous-time (continuous-time dynamics, continuous-time measurements) cases are analyzed. Also the corresponding optimal smoothers are analyzed.

In addition to the theory of optimal filtering, the purpose is also to present practical numerical algorithms for non-linear discrete-time, continuous-discrete time and continuous-time filtering and smoothing. As well as the classical extended Kalman filters and smoothers also the more recent unscented Kalman filter and smoother, particle filter and smoother, and the idea of Rao-Blackwellization of particle methods are analyzed. Unscented filters and smoothers are presented for the discrete-time, continuous-discrete and continuous-time filtering problems, and particle filters are presented for discrete-time and continuous-discrete time cases. The algorithms are also empirically compared using numerical simulations, which illustrate the performance of different algorithms while trying to imitate physical situations that might arise in real world applications.

Multiple target tracking problems, where an unknown number of moving targets is measured using multiple sensors is a very important application of the continuous-discrete filtering. The dynamics of the targets are most naturally modeled with stochastic differential equations and the continuous-discrete formulation is necessary, because the measurements arrive at irregular instances of time. Because of the multiple sensors the probabilistic treatment of the uncertainties in the system is the most appropriate. In this thesis, it is shown how particle filters can be used for estimating the states of the multiple targets and how Rao-Blackwellization can be used for improving the efficiency of the particle filter. The idea behind the Rao-Blackwellization is that by conditioning on the data associations and the birth and death processes, the posterior distributions of the states of the targets can be approximated with Gaussian distributions. For this reason, the target states can be integrated out analytically by the Kalman filter and the particle filter only needs to be applied to the data associations and to the birth and death processes. This significantly reduces the computational requirements and increases the efficiency of the particle filter.

Spread of an infectious disease in population can be modeled by a differential equation model, which is measured at discrete instances of time. If some of the model parameters are unknown, as often is the case, the on-line estimation of the spread and parameters can be formulated as a continuous-discrete filtering problem. Because of the strong non-linearities in both dynamic and measurement models, the filtering problem is most efficiently solved by particle filtering methods. The structure of the model also allows closed form marginalization, that is, Rao-Blackwellization of certain model parameters, which enhances the efficiency of the particle filter.

The theory of continuous-discrete Kalman filtering is related to the Gaussian process regression and to the more general non-parametric Bayesian modeling, especially in the case of time series data. For this reason Kalman filters and Kalman smoothers are very well suited to non-parametric and semi-parametric modeling and estimation of both long term and short term dependencies in time series. In this thesis, also the winning solution to the CATS (Competition on Artificial Time Series) time series prediction competition is presented. The competition was organized as a special session of the IJCNN (International Joint Conference on Neural Networks) 2004 conference. The Kalman smoother based method gave the best prediction and won the first price in the competition. Similar methods could be applied to other time series prediction or analysis problems including both trend (long term) and auto-correlative (short term) components.

1.1.1 Contributions

The contributions of the thesis can be summarized as follows:

- **The matrix form unscented transform:**

  A new *matrix form of the unscented transform* is presented. Using this matrix form, the *unscented Kalman filter (UKF) and the unscented Kalman smoother (UKS)* are presented in *matrix form*, which eases the mathematical treatment of the equations and provides new insight to connections with other filtering and smoothing methods.

- **Continuous-discrete and continuous-time unscented Kalman filtering and smoothing:**

  By taking the formal limit of the unscented Kalman filter and the unscented Kalman smoother when the prediction time step size goes to zero the *continuous-discrete unscented Kalman filter and smoother* are derived. By taking the formal limit, when both the prediction and measurement time step sizes go to zero the *unscented Kalman-Bucy filter*, which is a continuous-time version of the unscented Kalman filter, is derived. These derivations would not be possible, or at least would be very much harder without the matrix formulations of UKF and UKS.

- **Continuous-discrete sequential importance resampling:**

  Novel Girsanov theorem based methods for performing *continuous-discrete sequential importance resampling*, that is, *continuous-discrete particle filtering* are derived. Also a *Rao-Blackwellized continuous-discrete particle filter* is developed.

- **Applications of continuous-discrete filtering:**

  The complete applications of *multiple target tracking* in the case of *unknown number of targets*, *estimating the spread of an infectious disease* based on measured data and the winning solution to the *CATS time series prediction competition* are presented. Several smaller simulation examples that model real physical situations involving non-linear continuous-time dynamics are also presented.

1.1.2 Outline of the Thesis

This thesis starts with short presentation of the existing and potential applications, and the general ideas of discrete-time, continuous-discrete and continuous-time filtering. These topics are covered in Section 1.2.

Chapter 2 contains short reviews of stochastic differential equations and Bayesian inference. The purpose of these sections is to introduce the notation and terminology used in this thesis, because especially the notation tends to vary much between the mathematical and the applied probability literature. The notation is chosen to be closer to the notation of applied Bayesian inference and optimal filtering than to the notation of mathematical probability.

Section 3.1 reviews the classical theory of discrete-time filtering and smoothing in Bayesian point of view and brief derivations of the Bayesian non-linear filtering, Bayesian non-linear smoothing, Kalman filtering and Kalman smoothing equations are also given. The discrete-time extended Kalman filter, the unscented Kalman filter, particle filtering and Rao-Blackwellized particle filtering and smoothing are also presented in probabilistic terms. The main contributions of this section are the matrix form of the unscented Kalman filter and the matrix form of the unscented Kalman smoother.

Section 3.2 presents the theory and relevant algorithms of continuous-discrete filtering. The section starts with presentation of the classical theory of continuous-discrete filtering as sequential solving of certain partial differential equations and application of the Bayes' rule. Then the continuous-discrete Kalman filter and smoother, and the extended Kalman filter and smoother are presented, which are the classical algorithms of continuous-discrete filtering and smoothing. The main contributions of this section are the continuous-discrete unscented Kalman filter, the continuous-discrete unscented Kalman filter smoother, the Girsanov theorem based continuous-discrete sequential importance sampling and continuous-discrete Rao-Blackwellized sequential importance sampling.

Section 3.3 briefly presents the theory of continuous-time optimal filtering and smoothing. The most general optimal filtering equations are not given, but the continuous-time Kalman-Bucy and extended Kalman-Bucy filters are covered. The main contribution of this section is the continuous-time unscented Kalman-Bucy filter.

Chapter 4 presents applications of the optimal filtering algorithms. Complete solutions are presented for tracking of an unknown number of targets, estimating the spread of an infectious disease and to the CATS time series prediction competition.

1.2 What is Optimal Filtering?

Optimal filtering refers to the methodology that can be used for estimating the states of time varying systems, which are indirectly observed through noisy measurements. The *state* of the system refers to the dynamic variables such as position, velocities and accelerations or orientation and rotational motion parameters, which describe the physical state of the system. The *noise* in the measurements refers to a noise in the sense that the measurements are uncertain, that is, even if we knew the true system state the measurements would not be deterministic functions of the state, but would have certain distribution of possible values. The time evolution of the state is modeled as a dynamic system, which is perturbed by a certain *process noise*. This noise is used for modeling the uncertainties in the system dynamics and in most cases the system is not truly stochastic, but the stochasticity is only used for representing the model uncertainties.

1.2.1 Applications of Optimal Filtering

Phenomena, which can be modeled as time varying systems of the above type are very common in engineering applications. These kind of models can be found, for example, in navigation, aerospace engineering, space engineering, remote surveillance, telecommunications, physics, audio signal processing, control engineering, finance and several other fields. Examples of such applications are the following:

- *Global positioning system (GPS)* (Kaplan, 1996) is a widely used satellite navigation system, where the GPS receiver unit measures arrival times of signals from several GPS satellites and computes its position based on these measurements. The GPS receiver typically uses an extended Kalman filter or some other optimal filtering algorithm for computing the position and velocity such that the measurements and the assumed dynamics (laws of physics) are taken into account. Also the ephemeris information, which is the satellite reference information transmitted from the satellites to the GPS receivers is typically generated using optimal filters.

- *Target tracking* (Bar-Shalom et al., 2001) refers to the methodology, where a set of sensors such as active or passive radars, radio frequency sensors, acoustic arrays, infrared sensors and other types of sensors are used for determining the position and velocity of a remote target. When this tracking is done continuously, the dynamics of the target and measurements from the different sensors are most naturally combined using an optimal filter. The target in this (single) target tracking case can be, for example, a robot, a satellite, a car or an airplane.

- *Multiple target tracking* (Bar-Shalom and Li, 1995; Blackman and Popoli, 1999; Stone et al., 1999) systems are used for remote surveillance in the cases, where there are multiple targets moving at the same time in the same geographical area. This arises the concept of data association (which measurement was from which target?) and the problem of estimation of the number of targets. Multiple target tracking systems are typically used in remote surveillance for military purposes, but possible civil applications are, for example, monitoring of car tunnels, automatic alarm systems and people tracking in buildings.

- *Inertial navigation* (Titterton and Weston, 1997; Grewal et al., 2001) uses inertial sensors such as accelerometers and gyroscopes for computing the position and velocity of a device such as a car, an airplane or a missile. When the inaccuracies in sensor measurements are taken into account the natural way of computing the estimates is by using an optimal filter. Also in sensor calibration, which is typically done in time varying environment optimal filters are often applied.

- *Integrated inertial navigation* (Grewal et al., 2001; Bar-Shalom et al., 2001) combines the good sides of unbiased but inaccurate sensors, such as altimeters and landmark trackers, and biased but locally accurate inertial sensors. Combining of these different sources of information is most naturally performed using an optimal filter such as the extended Kalman filter. This kind of approach was used, for example, in the guidance system of Apollo 11 lunar module (Eagle), which landed on the moon in 1969.

- *GPS/INS navigation* (Grewal et al., 2001; Bar-Shalom et al., 2001) is a form of integrated inertial navigation, where the inertial sensors are combined with a GPS receiver unit. In GPS/INS navigation system the short term fluctuations of the GPS can be compensated with the inertial sensors and the inertial sensor biases can be compensated with the GPS receiver. An additional advantage of this approach is that it is possible to temporarily switch to pure inertial navigation, when the GPS receiver is unable to compute its position (i.e., has no fix) for some reason. This happens, for example, indoors, in tunnels and in other cases when there is no direct line-of-sight between the GPS receiver and the satellites.

- *Spread of infectious diseases* (Anderson and May, 1991; Hethcote, 2000) can often be modeled as differential equations for the number of susceptible, infected and recovered/dead individuals. When uncertainties are induced into the dynamic equations, and when the measurements are not perfect, the estimation of the spread of a disease can be formulated as an optimal filtering problem.

- *Biological processes* (Murray, 1993) such as population growth, predator-pray models and several other dynamic processes in biology can also be modeled as (stochastic) differential equations. The estimation of the states of these processes from inaccurate measurements can be formulated as an optimal filtering problem.

- *Telecommunications* is also a field where optimal filters are traditionally used. For example, optimal receivers, signal detectors and phase locked loops can be interpreted to contain optimal filters (Van Trees, 1968, 1971) as components. Also the celebrated Viterbi algorithm (Viterbi, 1967) can be interpreted as a combination of optimal filtering and optimal smoothing of the underlying hidden Markov model.

- *Audio signal processing* applications such as audio restoration (Godsill and Rayner, 1998) and audio signal enhancement (Fong et al., 2002) often use TVAR (time varying autoregressive) models as the underlying audio signal models. These kind of models can be efficiently estimated using optimal filters and smoothers.

- *Stochastic optimal control* (Maybeck, 1982b; Stengel, 1994) considers control of time varying stochastic systems. Stochastic controllers can typically be found in, for example, airplanes, cars and rockets. The optimality, in addition to the statistical optimality, means that control signal is constructed to minimize a performance cost, such as expected time to reach the target state, the amount of fuel consumed or average distance from a desired position trajectory. Optimal filters are typically used for estimating the states of the stochastic system and a deterministic optimal controller is constructed independently from the filter such that it uses the estimate of the filter as the known state. In theory the optimal controller and optimal filter are not completely decoupled and the problem of constructing optimal stochastic controllers is far more challenging than constructing optimal filters and (deterministic) optimal controllers separately.

- *Learning systems* or adaptive systems can often be mathematically formulated in terms of optimal filters. The theory of stochastic differential equations has close relationship with Bayesian non-parametric modeling, machine learning and neural network modeling (MacKay, 1998; Bishop, 1995). Methods, which are similar to the data association methods in multiple target tracking are also applicable to on-line adaptive classification (Andrieu et al., 2002).

- *Physical systems* which are time varying and measured through unideal sensors can sometimes be formulated as stochastic state space models, and the time evolution of the system can be estimated using optimal filters (Kaipio and Somersalo, 2005). In Vauhkonen (1997) and more recently, for example, in Pikkarainen (2005) optimal filtering is applied to Electrical Impedance Tomography (EIT) problem in time varying setting.

All the applications described above can be modeled as *discrete-time*, *continuous-discrete* or *continuous-time* optimal filtering problems, depending on whether the state and measurements are discrete or continuous functions of time. These concepts will be introduced next.

1.2.2 Origins of Bayesian Optimal Filtering

The roots of Bayesian analysis of time dependent behavior are in the optimal linear filtering. The idea of constructing mathematically optimal recursive estimators was first presented for linear systems due to their mathematical simplicity and the most natural optimality criterion in both mathematical and modeling point of view was the least squares optimality. For linear systems the optimal Bayesian solution (with MMSE utility) coincides with the least squares solution, that is, the optimal least squares solution is exactly the posterior mean.

The history of optimal filtering starts from the *Wiener filter* (Wiener, 1950), which is a spectral domain solution to the problem of (least squares) optimal filtering of stationary Gaussian signals. The Wiener filter is still important in communication applications (Van Trees, 1968) and digital signal processing (Hayes, 1996). The disadvantages of the Wiener filter are that it can only be applied to stationary signals and that the construction of a Wiener filter is often mathematically demanding and these mathematics cannot be avoided (i.e., made transparent). Due to the demanding mathematics the Wiener filter can only be applied to simple low dimensional filtering problems.

The success of optimal linear filtering in engineering applications is mostly due to the seminal article of Kalman (1960b), which describes the recursive solution to the optimal discrete-time (sampled) linear filtering problem. The reason to the success is that the *Kalman filter* can be understood and applied with very much lighter mathematical machinery than the Wiener filter. Also, despite its mathematical simplicity, the Kalman filter (or actually the Kalman-Bucy filter; Kalman and Bucy, 1961) contains the Wiener filter as its limiting special case.

In the early stages of its history, the Kalman filter was soon discovered to belong to the class of Bayesian estimators (Ho and Lee, 1964; Lee, 1964; Jazwinski, 1966, 1970). An interesting historical detail is that while Kalman and Bucy were formulating the linear theory in the United States, Stratonovich was doing the pioneering work on the probabilistic (Bayesian) approach in Russia (Stratonovich, 1968; Jazwinski, 1970).

As discussed in the book of West and Harrison (1997), in the sixties, Kalman filter like recursive estimators were also used in the Bayesian community and it is not clear whether the theory of Kalman filtering or the theory of *dynamic linear* *models* (DLM) was the first. Although these theories were originally derived from slightly different starting points, they are equivalent. Because of Kalman filter's useful connection to the theory and history of stochastic optimal control, this thesis approaches the Bayesian filtering problem from the Kalman filtering point of view.

Although the original derivation of the *Kalman filter* was based on the least squares approach, the same equations can be derived from the pure probabilistic Bayesian analysis. The Bayesian analysis of Kalman filtering is well covered in the classical book of Jazwinski (1970) and more recently in the book of Bar-Shalom et al. (2001). Kalman filtering, mostly because of its least squares interpretation, has widely been used in stochastic optimal control. A practical reason to this is that the inventor of the Kalman filter, Rudolph E. Kalman, has also made several contributions (Kalman, 1960a) to the theory of *linear quadratic Gaussian* (LQG) regulators, which are fundamental tools of stochastic optimal control (Stengel, 1994; Maybeck, 1982b).

1.2.3 Optimal Discrete-Time Filtering

Optimal discrete-time Bayesian filtering (see, e.g. Jazwinski, 1970; Bar-Shalom et al., 2001; Doucet et al., 2001; Ristic et al., 2004) considers statistical inversion problems, where the unknown quantity is a vector valued time series $(\mathbf{x}_1, \mathbf{x}_2, \ldots)$ which is observed through noisy measurements $(\mathbf{y}_1, \mathbf{y}_2, \ldots)$ as illustrated in the Figure 1.1. An example of this kind of time series is shown in the Figure 1.2. The process shown is actually a discrete-time noisy resonator with a known angular velocity. The state $\mathbf{x}_k = (x_k \; \dot{x}_k)^T$ is two dimensional and consists of the position of the resonator x_k and its time derivative $\dot{x}_k$. The measurements y_k are scalar observations of the resonator position (signal) and they are corrupted by measurement noise.

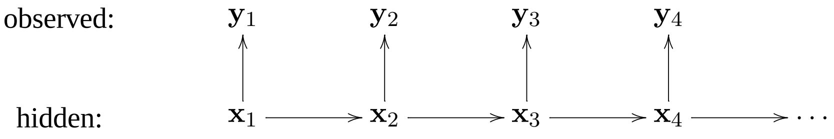

Figure 1.1: In discrete-time filtering a discrete sequence of hidden states $\mathbf{x}_k$ is indirectly observed through noisy measurements $\mathbf{y}_k$.

The purpose of the *statistical inversion* at hand is to estimate the hidden states $\{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$ given the observed measurements $\{\mathbf{y}_1, \ldots, \mathbf{y}_T\}$, which means that in the Bayesian sense (Bernardo and Smith, 1994; Gelman et al., 1995) the purpose is to compute the joint posterior distribution of all the states given all the

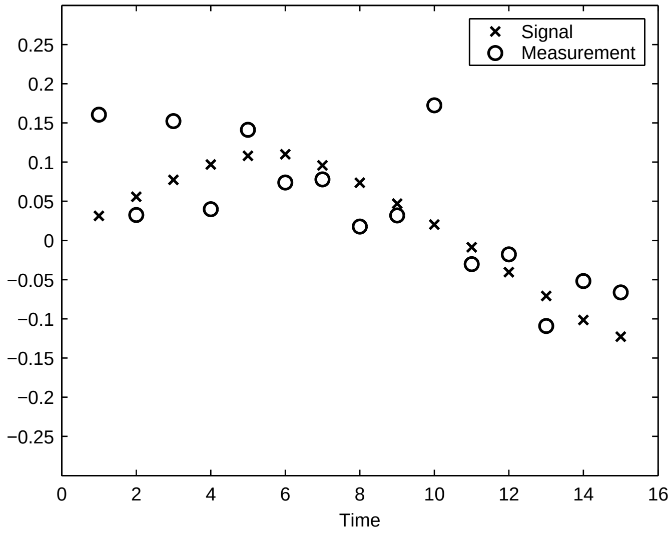

Figure 1.2: An example of time series, which models a discrete-time resonator. The actual resonator state (signal) is hidden and only observed through the noisy measurements.

measurements. In order to do that prior and likelihood models for the time series are needed (called the *filtering model*):

- **Initial distribution** specifies the *prior distribution* $p(\mathbf{x}_0)$ of the hidden state $\mathbf{x}_0$ at initial time step $k = 0$.
- **Dynamic model** models the system dynamics and its uncertainties as a *Markov sequence*, defined in terms of the transition distribution $p(\mathbf{x}_k \,|\, \mathbf{x}_{k-1})$.
- **Measurement model** models how the measurement $\mathbf{y}_k$ depends on the current state $\mathbf{x}_k$. This dependence is modeled by specifying the distribution of the measurement given the state $p(\mathbf{y}_k \,|\, \mathbf{x}_k)$.

Because computing the full joint distribution of the states at all time steps is computationally very inefficient and unnecessary in real-time applications, in *optimal (Bayesian) filtering* the objective is to compute the *filtering distributions*

$$p(\mathbf{x}_k \,|\, \mathbf{y}_1, \ldots, \mathbf{y}_k), \qquad k = 1, \ldots, T. \tag{1.1}$$

The filtering distribution is the marginal distribution of *the current state* $\mathbf{x}_k$ given *the previous measurements* $\{\mathbf{y}_1, \ldots, \mathbf{y}_k\}$. It turns out that these distributions can be computed *recursively* by the Bayesian filtering equations (Ho and Lee, 1964; Lee, 1964), which have the following form:

- **Prediction step** uses the *Chapman-Kolmogorov equation* for computing the step-a-head prediction distribution of the state $\mathbf{x}_k$.
- **Update step** uses the *Bayes' rule* for computing the posterior distribution of the state $\mathbf{x}_k$ given the current measurement $\mathbf{y}_k$.

The Bayesian filtering equations are *recursive,* which means that the previous computations do not need to be redone at each measurement and the amount of computations is, in principle, constant per time step. An example of filtering result is shown in the Figure 1.3. The prediction step can also be applied recursively to compute the *prediction distributions*

$$p(\mathbf{x}_{k+n} \,|\, \mathbf{y}_1, \ldots, \mathbf{y}_k), \qquad k = 1, \ldots, T, \quad n = 1, 2, \ldots, \tag{1.2}$$

which are the marginal distributions of the future states, n steps after the current time step.

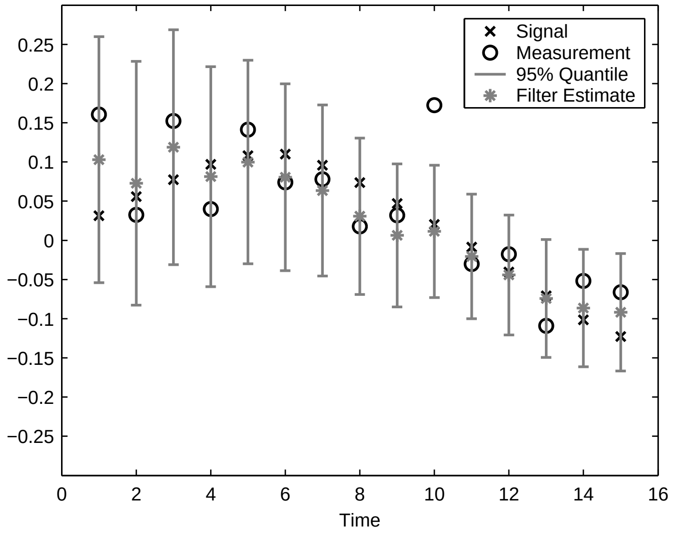

Figure 1.3: The result of computing the filtering distributions for the discrete-time resonator model. The *estimates* are the posterior means of the filtering distributions and the quantiles are the 95% quantiles of the filtering distributions.

Additional useful distributions are the *smoothing distributions*

$$p(\mathbf{x}_k \,|\, \mathbf{y}_1, \ldots, \mathbf{y}_T), \qquad k = 1, \ldots, T, \tag{1.3}$$

which can also be computed with recursive equations requiring a constant amount of computations per time step. These distributions are the marginal distributions of the states $\mathbf{x}_k$ given a certain interval $\{\mathbf{y}_1, \ldots, \mathbf{y}_T\}$ of measurements with $T > k$. The difference to the filtering distributions is that the smoothed distributions are conditional to the *future* (k > T) measurements also. An example of a smoothing result is shown in the Figure 1.4.

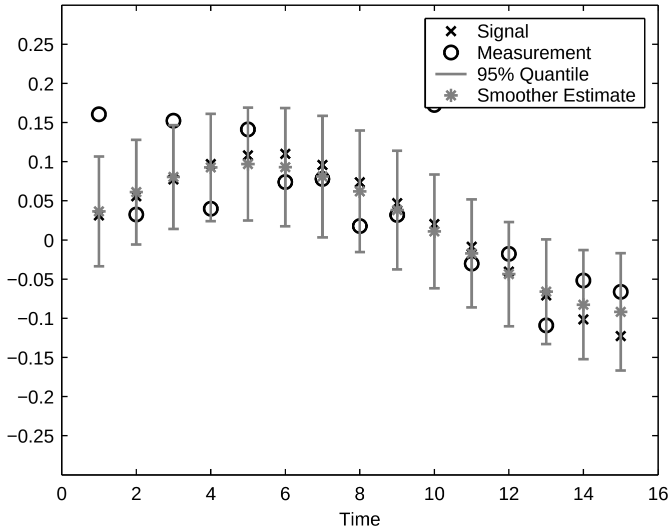

Figure 1.4: The result of computing the smoothing distributions for the discrete-time resonator model. The *estimates* are the posterior means of the smoothing distributions and the quantiles are the 95% quantiles of the smoothing distributions. The smoothing distributions are actually the marginal distributions of the full state posterior distribution.

Because the equations of Bayesian filtering can rarely be solved in closed form, a common approach is to approximate the non-linearities with linear functions or to explicitly approximate the filtering distributions by Gaussian distributions. The former approach is employed in the *extended Kalman filter* (see, e.g., Jazwinski, 1970; Maybeck, 1982a; Bar-Shalom et al., 2001; Grewal and Andrews, 2001), which is perhaps the most commonly used variation of the Kalman filter. The latter approach is used in the *unscented Kalman filter* (see, e.g., Julier et al., 1995; Julier and Uhlmann, 2004b; Wan and van der Merwe, 2001), where the *unscented transform* is used for approximating the evolution of Gaussian distribution in non-linear transforms.

Significantly different approach to optimal filtering is *particle filtering* (Gordon et al., 1993; Kitagawa, 1996), where a set of Monte Carlo samples is used for approximating the filtering solution. The idea of particle filtering is not new (see, e.g., Akashi and Kumamoto, 1977), but until recently there have not been computers that are powerful enough for real time computation of the particle filtering solutions. Particle filtering is well covered, for example, in the books of Doucet et al. (2001) and Ristic et al. (2004).

Finite state hidden Markov models (HMM) are also a class of models, which allow closed form solutions to the optimal filtering and smoothing equations (see, e.g., Ristic et al., 2004). In communications literature the optimal smoothing solution (actually the MAP estimate) to the HMM model is referred to as the Viterbi algorithm (Viterbi, 1967). In *grid based methods* the distributions are approximated by discrete distributions and the computations are performed in the similar manner as in hidden Markov models.

van der Merwe (2004) generalizes the unscented Kalman filters and other similar filters such as central difference filters to sigma point filters, which are a class of derivative-free non-linear Gaussian approximation based filters. Statistical linearization (Gelb, 1974) is related to the theory of describing functions and the resulting approximations are much similar to the approximations obtained by the unscented Kalman filter (van der Merwe, 2004).

Mixtures of Gaussian distributions can also be used for approximating the filtering and smoothing distributions. In Gaussian sum methods (Alspach and Sorenson, 1972) and in multiple model estimators (see, e.g., Bar-Shalom et al., 2001; Ristic et al., 2004) either the filtering distributions of the model are approximated as Gaussian mixtures or the model itself is formulated as a latent variable model. Generalized pseudo-Bayesian estimators (GPB) and interacting multiple model (IMM) estimators (see, e.g., Bar-Shalom et al., 2001) are extensions of the Gaussian mixture approximations, where the latent variables are modeled as a Markov chain.

In multiple target tracking context there are a number of methods, which can also be used in more general context. Multiple target tracking and the related estimation methods are discussed in Section 4.1.

The optimal smoothing methods have evolved at the same time as filtering methods, and as in the filtering case the optimal smoothing equations (Lee, 1964) can be solved in closed form only in a few special cases. The linear Gaussian case is such special case, and it leads to the Rauch-Tung-Striebel smoother (Rauch et al., 1965). Analogously to the extended Kalman filters there also exists extended Rauch-Tung-Striebel (or Kalman) smoothers (Sage and Melsa, 1971; Gelb, 1974), which use linearization for handling non-linear models. There also exists unscented Kalman smoother (Wan and van der Merwe, 2001), where the linearization of the extended Kalman smoother is replaced with the unscented transform.

Particle filters, in principle, produce smoothed estimates automatically and all one has to do is to store the samples from previous time steps (Kitagawa, 1996). However, this can lead to very degenerate approximations (Kitagawa, 1996; Doucet et al., 2000) and for this purpose better methods have been developed (Godsill et al., 2004).

1.2.4 Optimal Continuous-Discrete Filtering

Because in the Nature time is continuous, not discrete, often a physically more realistic approach than discrete-time filtering is *continuous-discrete filtering* (Jazwinski, 1966, 1970). In continuous-discrete filtering the state dynamics are modeled as *continuous*-time stochastic processes, that is, as *stochastic differential equations* (Karatzas and Shreve, 1991; Øksendal, 2003) and the measurements are obtained at *discrete* instances of time. This differs from the discrete-time filtering, because in that approach both the dynamics and measurements are modeled as discrete-time processes.

The idea of continuous-discrete filtering can be illustrated by considering a time series, which is not measured on each time step but instead between the discrete time steps (measurement steps) there are additional states as shown in the Figure 1.5. If we now imagine that we add an infinite number of additional states between the measurements, the state sequence becomes a *random function*, which is observed at discrete instances of time. This kind of process is visualized in the Figure 1.6. The process is a simulation of a discretely observed noisy continuous-time resonator.

$$\begin{tikzcd}
\text{observed:} & \mathbf{y}_1 & & & & \mathbf{y}_2 & & & \mathbf{y}_3 & \cdots \\
\text{hidden:} & \mathbf{x}(t_1) \arrow[u] \arrow[r] & \mathbf{x}(t_1') \arrow[r] & \mathbf{x}(t_1'') & \cdots & \mathbf{x}(t_2) \arrow[u] \arrow[r] & \mathbf{x}(t_2') & \cdots & \mathbf{x}(t_3) \arrow[u] & \cdots
\end{tikzcd}$$

Figure 1.5: Continuous-discrete filtering can be considered as limiting case of discrete-time filtering, where we add an infinite number of states between the measurements.

The random functions are mathematically modeled as *stochastic differential equations* (SDE), which can be considered as ordinary differential equations driven by random white noise processes $\mathbf{w}(t)$ as follows:[1]

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{f}(\mathbf{x}, t) + \mathbf{L}(t)\,\mathbf{w}(t). \tag{1.4}$$

Here $\mathbf{x}(t)$ is the state, $\mathbf{f}(\mathbf{x}, t)$ is the drift function, $\mathbf{L}(t)$ is the dispersion matrix, and $\mathbf{w}(t)$ is the white noise process.

In continuous-discrete filtering the measurements $\mathbf{y}_k$ are obtained at discrete time instances $\{t_1, t_2, ...\}$. The measurement model is of the same form as in

[1] Formally, stochastic differential equations are defined in terms of Itô integrals with respect to Brownian motions, but this issue is considered later in this thesis. This less rigorous white noise interpretation of the equations is more common in applied literature.

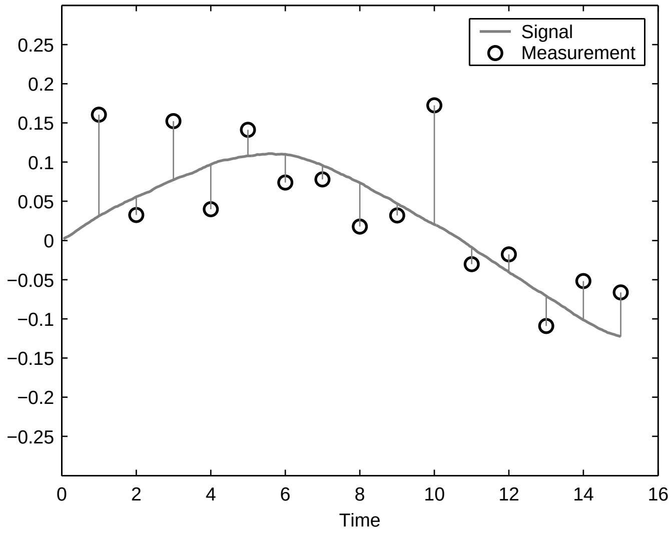

Figure 1.6: An example of continuous-discrete filtering data.

discrete-time filtering. However, to emphasize that the measurement model is a function of state at time t_k, not time index k, it is written as $p(\mathbf{y}_k \,|\, \mathbf{x}(t_k))$.

The advantage of the continuous-discrete model formulation over the discrete model formulation is that the time step size $\Delta t_k = t_{k+1} - t_k$ does not need to be constant, that is, the sampling interval may change as function of time.

In theory, the optimal (Bayesian) solution to the continuous-discrete filtering problem can be computed by the following prediction and update steps (Jazwinski, 1966, 1970):

- **Prediction step** solves the predicted probability density at time step t_k from the *Kolmogorov forward partial differential equation* using the old posterior probability density at time step t_{k-1} as the boundary condition.

- **Update step** uses the *Bayes' rule* for computing the posterior probability density of state at time step t_k from the predicted probability density of the prediction step, and the measurement $\mathbf{y}_k$.

An example of continuous-discrete filtering is shown in the Figure 1.7. In the figure the characteristic property of the continuous-discrete filtering solution can be seen: the filtering estimate changes smoothly between the measurements and jumps (has discontinuity) at the measurements. This is because the information in the measurements is discrete in nature, but between the measurements only information from the continuous-time dynamic model is used.

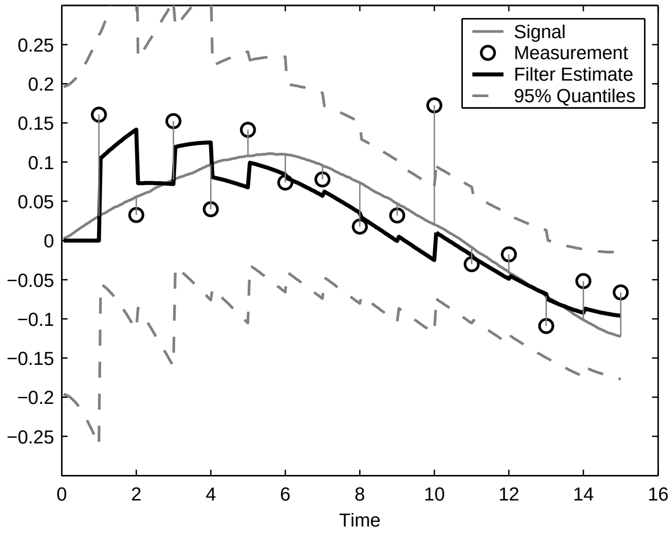

Figure 1.7: An example of continuous-discrete filtering result.

The continuous-discrete smoothing solution (Leondes et al., 1970) can be computed for each instance of time and the smoothing solution is typically a continuous function of time. An example of smoothing solution is shown in the Figure 1.8.

The *filtering, smoothing and prediction distributions* are of the following form:

- **Filtering distribution** is the distribution of the state $\mathbf{x}(t_k)$ at the time instance t_k, which is the time when the measurement $\mathbf{y}_k$ has been obtained:

  $$p(\mathbf{x}(t_k) \,|\, \mathbf{y}_1, \ldots, \mathbf{y}_k). \tag{1.5}$$

  The filtering distributions can be extended to be defined for all t by replacing the filtering distribution with the prediction distribution between the measurements.

- **Prediction distributions** can be computed for all times $t > t_k$

  $$p(\mathbf{x}(t) \,|\, \mathbf{y}_1, \ldots, \mathbf{y}_k), \qquad t > t_k \tag{1.6}$$

  by solving the corresponding Kolmogorov forward equation. The marginal posterior distributions of the states between the measurements conditional to all the previous measurements are given by these prediction distributions.

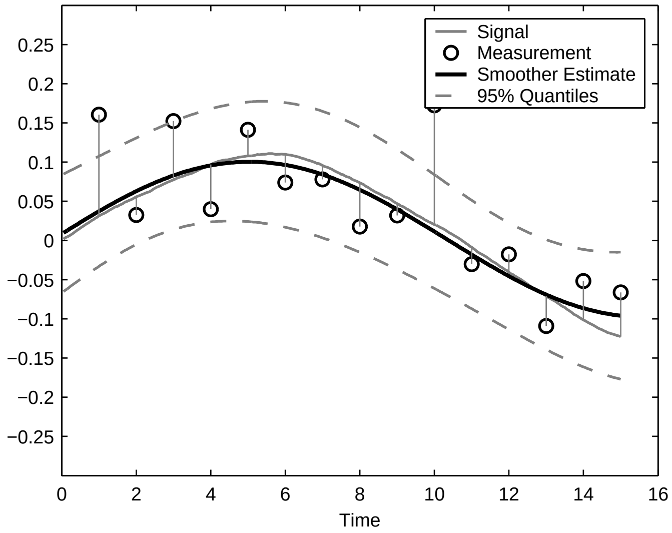

Figure 1.8: Example of continuous-discrete smoothing result.

- **Smoothing distributions** can be computed for all times $t \in [0, t_T]$ if the measurements up to the time instance $\mathbf{y}_T$ have been observed:

$$p(\mathbf{x}(t) \,|\, \mathbf{y}_1, \ldots, \mathbf{y}_T), \qquad 0 < t < t_T. \tag{1.7}$$

Thus the smoothing distributions are also defined continuously for all t.

Instead of solving the corresponding Kolmogorov forward equation it is also possible to find a (weak or strong) stochastic process solution to the stochastic differential equation directly. In theory, it is always possible to find at least a weak solution such that it can be written in discrete-time form with a suitably defined transition density $p(\mathbf{x}(t_k) \,|\, \mathbf{x}(t_{k-1}))$. This kind of discrete-time process is a solution in the sense that the state distributions at measurement times t_k are the same as the state distributions of the continuous-time process. Finding this kind of *equivalent discrete-time model* is often extremely difficult, because it requires a closed form solution to the (law of) stochastic differential equation. In most cases the corresponding discrete time transition density is not Gaussian, and often it is infinite-dimensional or cannot be even represented in terms of elementary functions.

The idea of continuous-discrete filtering itself is not new and the classic book about the Bayesian approach to continuous-discrete filtering problems is the book of Jazwinski (1970). Actually, in the seminal article (Kalman, 1960b) also the discretization of continuous-time dynamic systems is discussed and in that sense it already covers the continuous-discrete Kalman filter, not only the discrete-time Kalman filter.

Because trying to find the closed form expression of the law of stochastic process in terms of, for example, Brownian motion and its integrals is very difficult in general and can be practically applied only in a few special cases (Kalman, 1960b; Beneš, 1981; Daum, 1984, 1986), a more feasible approach is to find the solution numerically. The continuous-discrete extended Kalman filter (EKF) (Jazwinski, 1970; Gelb, 1974; Grewal and Andrews, 2001) uses a Taylor series approximation to the non-linear drift function $\mathbf{f}(\cdot)$ and forms a Gaussian process approximation to the SDE. Another possible approach is to simulate sample paths of the SDE (Kloeden and Platen, 1999) and use particle filters for estimation (Doucet et al., 2001; Ristic et al., 2004). Continuous-discrete filtering has been well developed in context of Kalman filters and extended Kalman filters, but a continuous-discrete version of the unscented Kalman filter has not existed up to this point. Also there has not existed continuous-discrete particle filters significantly beyond the bootstrap filter.

In Section 3.2 of this thesis it is shown how the solution of the SDE can be approximated with a Gaussian process by using a continuous-time form of the unscented Kalman filter (Julier and Uhlmann, 2004b; Julier et al., 1995; Wan and van der Merwe, 2001). In the same section also novel measure transformation based methods to continuous-discrete sequential importance resampling are presented. The methods are based on transformations of probability measures by the Girsanov theorem. Measure transformation based approaches are particularly successful in continuous-time filtering (Kallianpur, 1980), but have less been used in continuous-discrete filtering.

The general idea of using the Girsanov theorem in importance sampling of SDEs has been presented, for example, in Kloeden and Platen (1999). Ionides (2004) presents idea of using transformations of probability measures for computing the likelihood ratios between importance process and the true process in context of continuous-discrete filtering. However, the results of Ionides (2004) only apply when the Euler integration scheme is used and when the dispersion matrix is invertible.

Interacting and branching particle systems are particle based solutions to non-linear filtering problems also in the continuous-discrete setting. In these methods the Girsanov theorem is used for transforming the measure of the observation process. A thorough review of these methods can be found in (Moral and Miclo, 2000). The convergence rates of these methods have been studied recently in (Kouritzin and Sun, 2005).

Roberts and Stramer (2001) and Elerian et al. (2001) present MCMC (Markov chain Monte Carlo) based methods for computing the full posteriors (smoothing solutions) of scalar diffusion processes, which are observed at discrete instances of time. The methods are based on modeling the paths between the data points as missing data. However, they are not filtering algorithms, because they are batch (non-recursive) algorithms and the states itself are measured at the discrete time steps without an error. Eraker (2001) presents a similar method for multidimensional case, and the model is more general in the sense that the state is allowed to be only partially measured, as in filtering models. Unfortunately, the method can only be applied to the case where diffusion matrix of the state process is invertible, which rules out many physically relevant models.

In the non-linear projection filter (Gunther et al., 1997) the idea is to integrate the Kolmogorov forward equation of the non-linear dynamic model by the method of Galerkin (see, e.g., Guenther and Lee, 1988). In (Challa et al., 2000) the solution of the Kolmogorov equation is approximated with generalized Edgeworth series and Gauss-Hermite quadrature. In Gaussian sum filter (Alspach and Sorenson, 1972) the posterior density is approximated using a weighted sum of Gaussian distributions. Statistical linearization (Gelb, 1974) can also be used in the continuous-discrete filtering case. Generalized pseudo-Bayesian estimators (GPB) and interacting multiple model (IMM) estimators (see, e.g., Bar-Shalom et al., 2001) as well as multiple model estimators (see, e.g., Bar-Shalom et al., 2001; Ristic et al., 2004) are also applicable in the continuous-discrete filtering case.

In *multiple target tracking* (see, Section 4.1) the sensors are typically asynchronous and for this reason most multiple target tracking methods are based on continuous-discrete filters. Also the dynamics of the targets are often most naturally modeled in continuous-time, which leads to stochastic differential equations that are observed at discrete instances of time.

Optimal continuous-discrete smoothing can be performed using the discrete-time smoothing equations once the transition densities have been solved and if the smoothing solution at the measurement times is enough. When also the times between the measurements are considered the continuous-discrete smoothing is closer to continuous-time smoothing than to discrete-time smoothing. The partial differential equations of the continuous-discrete smoothing solution are given in Leondes et al. (1970).

In linear Gaussian case the continuous-time Rauch-Tung-Striebel smoother (Rauch et al., 1965) can be used in continuous-discrete models as such. Also the continuous-time extended Rauch-Tung-Striebel (or Kalman) smoothers (Sage and Melsa, 1971; Gelb, 1974) are applicable in the continuous-discrete case.

In Section 3.2 of this thesis it is shown how the linearization of the continuous-time extended Kalman smoother can be replaced by the unscented transform, which leads to the novel continuous-time unscented Kalman smoother, which is also applicable in the continuous-discrete case. As in the discrete-time case, continuous-discrete particle filters also automatically produce an estimate of the smoothing solution, but the problems of this direct solution are the same as in discrete-time case. This issue is discussed in Section 3.2.

1.2.5 Optimal Continuous-Time Filtering

$$\begin{tikzcd}
\text{observed:} & \mathbf{y}(t) & \mathbf{y}(t+\mathrm{d}t) & \mathbf{y}(t+2\mathrm{d}t) & \mathbf{y}(t+3\mathrm{d}t) & \\
\text{hidden:} & \mathbf{x}(t) \arrow[u] \arrow[r] & \mathbf{x}(t+\mathrm{d}t) \arrow[u] \arrow[r] & \mathbf{x}(t+2\mathrm{d}t) \arrow[u] \arrow[r] & \mathbf{x}(t+3\mathrm{d}t) \arrow[u] \arrow[r] & \cdots
\end{tikzcd}$$

Figure 1.9: Continuous-time filtering can be considered as limiting case of discrete-time filtering, where the time interval between the successive states and the corresponding measurements goes to zero.

In analog communication systems the measured signals are typically continuous-time processes and analog receivers are devices, which demodulate or estimate the actual transmitted continuous-time signals from the noisy measured signals. Also in many analog (electrical or mechanical) control systems operating without digital computers the measured signals are continuous-time, not discrete-time signals.

Optimal continuous-time filtering (Bucy and Joseph, 1968; Jazwinski, 1970; Gelb, 1974; Liptser and Shiryayev, 1977; Kallianpur, 1980) considers stochastic inference problems, where the system model consists of a continuous-time state process $\mathbf{x}(t)$ and a continuous-time measurement process $\mathbf{z}(t)$. Thus, continuous-time filtering considers the limiting filtering problem, where the time interval between the states and measurements goes to zero, as illustrated in the Figure 1.9. In the limit $\mathrm{d}t \to 0$ both the state and measurement sequences become random functions, which can be modeled as stochastic differential equations of the form[2]

$$\begin{aligned}
\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} &= \mathbf{f}(\mathbf{x}(t), t) + \mathbf{L}(t)\,\mathbf{w}(t) \\
\mathbf{z}(t) &= \mathbf{h}(\mathbf{x}(t), t) + \mathbf{V}(t)\,\mathbf{e}(t),
\end{aligned} \tag{1.8}$$

where $\mathbf{w}(t)$ and $\mathbf{e}(t)$ are white noise processes.

The purpose of the continuous-time optimal (Bayesian) filter is to compute the posterior distribution (or the filtering distribution) of the process $\mathbf{x}(t)$ given the observed process $\{\mathbf{z}(\tau) \ : \ 0 \le \tau \le t\}$

$$p(\mathbf{x}(t) \,|\, \{\mathbf{z}(\tau) \ : \ 0 \le \tau \le t\}). \tag{1.9}$$

[2] As in the continuous-discrete filtering case, formally, the continuous-time filtering problem is defined in terms of Brownian motion driven Itô processes. Due to this the formal measurement process is actually the integral of measurement process $\mathbf{y}(t) = \int_0^t \mathbf{z}(t)\,\mathrm{d}t$. In this thesis we only consider non-linear continuous-time optimal filtering problems with additive noise, which is also the reason why we can safely use the white noise interpretation of the equations.

An example of continuous-time filtering is shown in the Figure 1.10. The underlying signal is the same noisy resonator as in the continuous-discrete filtering case in the previous section. The measurement signal $z(t)$ is a continuous-time process and it has been truncated a bit for visualization purposes such that it does not take values from minus infinity to plus infinity as it formally should.

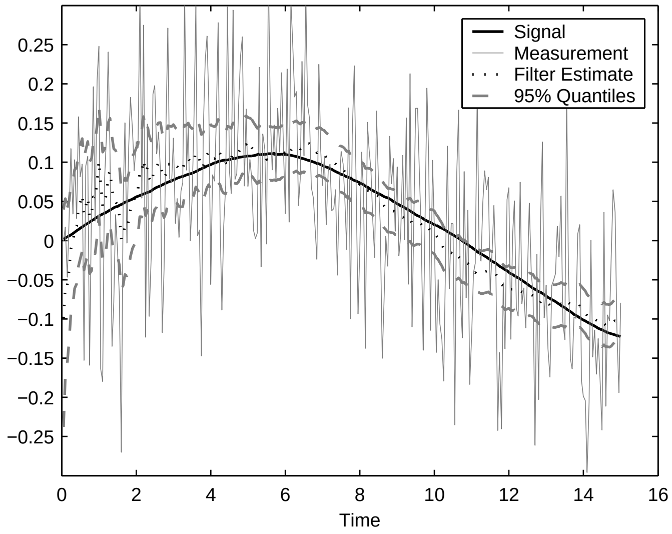

Figure 1.10: An example of continuous-time filtering result.

The optimal continuous-time smoothing equations (Leondes et al., 1970) can be used for computing the (smoothed) posterior distribution

$$p(\mathbf{x}(t) \,|\, \{\mathbf{z}(\tau) \;:\; 0 \leq \tau \leq T\}), \tag{1.10}$$

of the state at time t given the history of the measurements up to the time instance $T > t$.

The Kalman filter equations can be generalized to continuous-time measurement processes and the resulting filter is called the *Kalman-Bucy filter* (Kalman and Bucy, 1961). In this filter both the state process and measurements are modeled as linear stochastic differential equations. The continuous-time linear filtering equations consist of a vectorial linear differential equation for the mean and a non-linear matrix Riccati differential equation for the covariance.

The non-linear Bayesian filtering equations can also be generalized to continuous time and the resulting equation is called the *Kushner-Stratonovich equation* (Kushner, 1964; Stratonovich, 1968; Bucy, 1965), which is a measure valued stochastic partial differential equation. The unnormalized version of the equation is called *Zakai equation* (Zakai, 1969) (see also Kallianpur, 1980). However, the problem in these equations is that they only give the formal solution, and the actual computation of the distribution or its expectations would require an infinite amount of computational resources. For this reason, in general, approximations must be used. In certain cases, the equations do have finite dimensional solutions, which lead to Kalman-Bucy filters (Kalman and Bucy, 1961) and Beneš filters (Beneš, 1981).

The extended Kalman-Bucy filter (EKBF) (see, e.g., Gelb, 1974) approximates the exact solution by replacing the non-linear model with a suitably linearized approximate model, which can be solved by the Kalman-Bucy filter. The EKBF can be interpreted as a method for forming a Gaussian process approximation to the optimal filtering solution (i.e., the posterior process). In Section 3.3 the equations of the new unscented Kalman-Bucy filter are derived, in which the continuous-time optimal filtering solution is approximated by using the continuous-time form of the unscented transform. Another general way of forming approximations is Monte Carlo sampling (Crisan and Lyons, 1997, 1999), where a set of weighted particles is used for approximating the posterior probability measure.

More information on practical continuous-time Kalman filtering and linearization based non-linear filtering can be found in books of Gelb (1974), Stengel (1994), Bar-Shalom et al. (2001), and Grewal and Andrews (2001). A bit more theoretical material and also material on more general non-linear filtering can be found in books of Bucy and Joseph (1968) and Jazwinski (1970). Analysis of continuous-time filtering problems in Itô calculus point of view can be found, for example, in books of Liptser and Shiryayev (1977), Kallianpur (1980) and Øksendal (2003). Good introduction to this point of view is given in the lecture notes of Karatzas (1988).

For general treatment of the probabilistic continuous-time smoothing problem and the related stochastic partial differential equations, see (Liptser and Shiryayev, 1977). There also exists so called Zakai forms of these equations (Antonelli and Elliott, 1986; Krishnamurthy and Elliott, 2002). Leondes et al. (1970) presents partial differential equations, which can be used for computing the smoothing solution once the filtering solution has been computed.

As in the filtering case, the optimal continuous smoothing problem can be solved exactly only if the model is linear Gaussian (Rauch et al., 1965; Gelb, 1974). Approximate solutions to more general non-linear continuous optimal smoothing problems can be computed by using the extended continuous-time Kalman-Bucy (Rauch-Tung-Striebel) smoother (see, e.g., Leondes et al., 1970; Sage and Melsa, 1971; Gelb, 1974), which uses a linear or quadratic approximation of the non-linear continuous-time model. The novel continuous-discrete unscented Kalman smoother derived in Section 3.2 can also be used for approximating the smoothing solutions of continuous-time filtering models.

Chapter 2

Probabilistic Inference and Stochastic Processes

This chapter briefly presents the principles of Bayesian inference and the theory of stochastic differential equations. The mathematical complexity is kept to the minimum but still at the level, which should be enough for understanding the theory of optimal filtering in the next chapter.

2.1 Bayesian Inference

This section provides a brief presentation of the philosophical and mathematical foundations of Bayesian inference. The connections to the classical statistical inference are also briefly discussed.

2.1.1 Philosophy of Bayesian Inference

The purpose of Bayesian inference (Bernardo and Smith, 1994; Gelman et al., 1995) is to provide a mathematical machinery that can be used for modeling systems, where the uncertainties of the system are taken into account and the decisions are made according to rational principles. The tools of this machinery are the probability distributions and the rules of probability calculus.

If we compare the frequentist statistical analysis to Bayesian inference the difference is that in Bayesian inference the probability of an event does not mean the proportion of the event in an infinite number of trials, but the uncertainty of the event in a single trial. Because models in Bayesian inference are formulated in terms of probability distributions, the probability axioms and computation rules of the probability theory (see, e.g., Shiryaev, 1996) also apply in the Bayesian inference.

2.1.2 Connection to Frequentist Statistics

Consider a situation, where we know the functional form $p(\mathbf{y}_k \,|\, \mathbf{x})$ of the distribution of conditionally independent random variables (measurements) $\mathbf{y}_1, \ldots, \mathbf{y}_n$, but the parameter $\mathbf{x} \in \mathbb{R}^n$ is unknown. The classical statistical method for estimating the parameter is the *maximum likelihood method* (Milton and Arnold, 1995), where we maximize the joint probability of the measurements, also called the likelihood function

$$L(\mathbf{x}) = \prod_k p(\mathbf{y}_k \,|\, \mathbf{x}). \tag{2.1}$$

The maximum of the likelihood function with respect to $\mathbf{x}$ gives the *ML-estimate* (ML-estimate)

$$\hat{\mathbf{x}} = \arg\max_{\mathbf{x}} L(\mathbf{x}). \tag{2.2}$$

The difference between the Bayesian inference and the frequentist statistics is that the starting point of Bayesian inference is to consider the parameter $\mathbf{x}$ as a random variable. The posterior distribution of the parameter $\mathbf{x}$ can be computed by using the *Bayes' rule*

$$p(\mathbf{x} \,|\, \mathbf{y}_1, \ldots, \mathbf{y}_n) = \frac{p(\mathbf{y}_1, \ldots, \mathbf{y}_n \,|\, \mathbf{x})\, p(\mathbf{x})}{p(\mathbf{y}_1, \ldots, \mathbf{y}_n)}, \tag{2.3}$$

where $p(\mathbf{x})$ is the prior distribution, which models the prior beliefs of the parameter before we have seen any data and $p(\mathbf{y}_1, \ldots, \mathbf{y}_n)$ is a normalization term, which is independent of the parameter $\mathbf{x}$. Often this normalization constant is left out and if the measurements $\mathbf{y}_1, \ldots, \mathbf{y}_n$ are conditionally independent given $\mathbf{x}$, the posterior distribution of the parameter can be written as

$$p(\mathbf{x} \,|\, \mathbf{y}_1, \ldots, \mathbf{y}_n) \propto p(\mathbf{x}) \prod_k p(\mathbf{y}_k \,|\, \mathbf{x}). \tag{2.4}$$

Because we are dealing with a distribution, we might now choose the most probable value of the random variable (MAP-estimate), which is given by the maximum of the posterior distribution. However, better estimate in mean squared sense is the posterior mean of the parameter (MMSE-estimate). There are an infinite number of other ways of choosing the point estimate from the distribution and the best way depends on the assumed loss function (or utility function). The ML-estimate can be considered as a MAP-estimate with uniform prior on the parameter $\mathbf{x}$.

2.1.3 The Building Blocks of Bayesian Models

The basic blocks of a Bayesian model are the *prior model* containing the preliminary information on the parameter and the *likelihood model* determining the stochastic mapping from the parameter to the measurements. Using the combination rules, namely the Bayes' rule, it is possible to infer an estimate of the parameters from the measurements. The distribution of the parameters, which is conditional to the observed measurements is called the *posterior distribution* and it is the distribution representing the state of knowledge about the parameters when all the information in the observed measurements and the model is used. *Predictive posterior distribution* is the distribution of new (not yet observed) measurements when all the information in the observed measurements and the model is used.

- **Prior model**
  The prior information consists of subjective experience based beliefs on the possible and impossible parameter values and their relative likelihoods before anything has been observed. The prior distribution is a mathematical representation of this information:
  $$p(\mathbf{x}) = \text{Information on parameter } \mathbf{x} \text{ before seeing any observations.} \tag{2.5}$$
  The lack of prior information can be expressed by using a non-informative prior. The non-informative prior distribution can be selected in various different ways (Gelman et al., 1995).

- **Likelihood model**
  Between the true parameters and the measurements there often is a causal, but inaccurate or noisy relationship. This relationship is mathematically modeled using the likelihood distribution:
  $$p(\mathbf{y}_k \,|\, \mathbf{x}) = \text{Distribution of observations given the parameters } \mathbf{x}. \tag{2.6}$$

- **Posterior distribution**
  Posterior distribution is the distribution of the parameters after the measurement $\mathbf{y}$ has been obtained and it can be computed by using the Bayes' rule:
  $$p(\mathbf{x} \,|\, \mathbf{y}) = \frac{p(\mathbf{y} \,|\, \mathbf{x})\, p(\mathbf{x})}{p(\mathbf{y})}, \tag{2.7}$$
  where
  $$p(\mathbf{y}) = \int_{\mathbf{x}} p(\mathbf{y} \,|\, \mathbf{x})\, p(\mathbf{x})\, \mathrm{d}\mathbf{x}. \tag{2.8}$$
  In the case of multiple measurements $\mathbf{y}_1, \ldots, \mathbf{y}_n$, if the measurements are conditionally independent the joint likelihood of all measurements is the product of individual measurements and the posterior distribution is
  $$p(\mathbf{x} \,|\, \mathbf{y}_1, \ldots, \mathbf{y}_n) \propto p(\mathbf{x}) \prod_k p(\mathbf{y}_k \,|\, \mathbf{x}), \tag{2.9}$$
  where the normalization term can be computed by integrating the right hand side over $\mathbf{x}$. If the random variable is discrete the integration reduces to summation.

- **Predictive posterior distribution**
  The predictive posterior distribution is the distribution of new measurements $\mathbf{y}_{n+1}$:

$$p(\mathbf{y}_{n+1} \,|\, \mathbf{y}_1, \dots, \mathbf{y}_n) = \int_{\mathbf{x}} p(\mathbf{y}_{n+1} \,|\, \mathbf{x})\, p(\mathbf{x} \,|\, \mathbf{y}_1, \dots, \mathbf{y}_n)\, \mathrm{d}\mathbf{x}. \qquad (2.10)$$

  After obtaining the measurements $\mathbf{y}_1, \dots, \mathbf{y}_n$ the predictive posterior distribution can be used for computing the probability distribution for $n + 1$:th measurement, which has not been observed yet.

In the case of tracking, we could imagine that the parameter is the sequence of dynamic states of a target, where the state contains the position and velocity. Or in the continuous-discrete setting the parameter would be an infinite-dimensional random function describing the trajectory of the target at a given time interval. In both cases the measurements could be, for example, noisy distance and direction measurements produced by a radar.

2.1.4 Bayesian Point Estimates

The distributions as such have no use in applications, but also in Bayesian computations finite dimensional summaries (point estimates) are needed. This selection of a point from space based on observed values of random variables is a statistical decision, and therefore this selection procedure is most naturally formulated in terms of *statistical decision theory* (Berger, 1985; Bernardo and Smith, 1994; Raiffa and Schlaifer, 2000).

Definition 2.1 (Loss Function)**.** *A loss function $L(\mathbf{x}, \mathbf{a})$ is a scalar valued function, which determines the loss of taking the* action $\mathbf{a}$, *when the true parameter value is $\mathbf{x}$. The action (or control) is the statistical decision to be made based on the currently available information.*

Instead of loss functions it is also possible to work with utility functions $U(\mathbf{x}, \mathbf{a})$, which determine the reward from taking the action $\mathbf{a}$ with parameter values $\mathbf{x}$. Loss functions can be converted to utility functions and vice versa by defining $U(\mathbf{x}, \mathbf{a}) = -L(\mathbf{x}, \mathbf{a})$.

If the value of parameter $\mathbf{x}$ is not known, but the knowledge on the parameter can be expressed in terms of the posterior distribution $p(\mathbf{x} \,|\, \mathbf{y}_1, \dots, \mathbf{y}_n)$, then the natural choice is the action, which gives the *minimum (maximum) of the expected loss (utility)* (Berger, 1985):

$$\mathrm{E}[L(\mathbf{x}, \mathbf{a}) \,|\, \mathbf{y}_1, \dots, \mathbf{y}_n] = \int L(\mathbf{x}, \mathbf{a})\, p(\mathbf{x} \,|\, \mathbf{y}_1, \dots, \mathbf{y}_n)\, \mathrm{d}\mathbf{x}. \qquad (2.11)$$

Commonly used loss functions are the following:

- *Quadratic error loss*: If the loss function is quadratic

  $$L(\mathbf{x},\mathbf{a}) = (\mathbf{x}-\mathbf{a})^T(\mathbf{x}-\mathbf{a}), \tag{2.12}$$

  then the optimal choice $\mathbf{a}_o$ is the *posterior mean* of the distribution of $\mathbf{x}$:

  $$\mathbf{a}_o = \int \mathbf{x}\, p(\mathbf{x}\,|\,\mathbf{y}_1,\ldots,\mathbf{y}_n)\,\mathrm{d}\mathbf{x}. \tag{2.13}$$

  This posterior mean based estimate is often called the *minimum mean squared error (MMSE)* estimate of the parameter $\mathbf{x}$. The quadratic loss is the most commonly used loss function, because it is easy to handle mathematically and because in the case of Gaussian posterior distribution the MAP estimate and the median coincide with the posterior mean.

- *Absolute error loss*: The loss function of the form

  $$L(\mathbf{x},\mathbf{a}) = \sum_i |x_i - a_i|, \tag{2.14}$$

  is called an absolute error loss and in this case the optimal choice is the *median* of the distribution (i.e., medians of the marginal distributions in multidimensional case).

- *0-1 loss*: If the loss function is of the form

  $$L(\mathbf{x},\mathbf{a}) = \begin{cases} 1 & , \text{ if } \mathbf{x} = \mathbf{a} \\ 0 & , \text{ if } \mathbf{x} \neq \mathbf{a} \end{cases} \tag{2.15}$$

  then the optimal choice is the maximum of the posterior distribution, that is, the *maximum a posterior (MAP)* estimate of the parameter.

In the case of continuous-discrete filtering, the loss function can be interpreted to be a functional of the form

$$L(\mathbf{x},\mathbf{a}) = \int_0^{t_T} (\mathbf{x}(t)-\mathbf{a}(t))^T\,(\mathbf{x}(t)-\mathbf{a}(t))\;\mathrm{d}t, \tag{2.16}$$

and the mean estimate produced by the optimal smoother is the function $\mathbf{a}_o(t)$ such that the expected value of the loss function above is minimized.

Loss functions are also used as the performance criteria in stochastic optimal control (Maybeck, 1982b; Stengel, 1994). For example, in stochastic quadratic regulator problems the loss function at time t is typically of the form

$$L(\mathbf{x},\mathbf{a}) = \int_t^{t_T} \left[\mathbf{a}^T(t)\,\mathbf{U}\,\mathbf{a}(t) + \mathbf{x}^T(t)\,\mathbf{X}\,\mathbf{x}(t)\right]\;\mathrm{d}t, \tag{2.17}$$

where $\mathbf{U}$ and $\mathbf{X}$ are some positive definite matrices. However, stochastic control problems are much more complicated than simple expected loss minimization, because at every time instance there exists past measurements and future measurements, which have not been observed yet. Also the future measurements have to be modeled, because the optimal control (action) depends on how much information the future measurements give, which in turn depends on the future states that are indirectly determined by the selected control. This property of stochastic control is called the dual effect or the active learning property (Maybeck, 1982b).

Expected utilities and losses are also commonly used in financial applications (Karatzas and Shreve, 1991). Expected utilities have been also successfully applied to practical Bayesian model selection problems (Vehtari, 2001; Vehtari and Lampinen, 2003).

2.1.5 Numerical Methods

In principle, Bayesian inference provides the optimal way of solving any model once the model specification has been set up. However, the practical problem is that computation of the integrals involved in the equations can rarely be performed analytically and numerical methods are needed. The integrals tend to be over spaces with high dimensionality, which renders all discretization based integration method useless. Here we shall briefly describe numerical methods, which are also applicable in higher dimensional problems: Gaussian approximations, Monte Carlo methods and importance sampling.

- Very common types of approximations are *Gaussian approximations* (Gelman et al., 1995), where the posterior distribution is approximated with a Gaussian distribution

$$p(\mathbf{x} \,|\, \mathbf{y}_1, \ldots, \mathbf{y}_n) \approx \mathrm{N}(\mathbf{x} \,|\, \mathbf{m}, \mathbf{P}). \tag{2.18}$$

  The mean $\mathbf{m}$ and covariance $\mathbf{P}$ of the Gaussian approximation can be either computed by matching the first two moments of the posterior distribution, or by using the maximum of the distribution as the mean estimate and approximating the covariance with the curvature of the posterior on the mode.

- In direct *Monte Carlo methods* a set of N samples from the posterior distribution is drawn

$$\mathbf{x}^{(i)} \sim p(\mathbf{x} \,|\, \mathbf{y}_1, \ldots, \mathbf{y}_n), \qquad i = 1, \ldots, N, \tag{2.19}$$

  and expectation of any function $\mathbf{g}(\cdot)$ can be then approximated as the sample average

$$\mathrm{E}[\mathbf{g}(\mathbf{x}) \,|\, \mathbf{y}_1, \ldots, \mathbf{y}_n] \approx \frac{1}{N} \sum_i \mathbf{g}(\mathbf{x}^{(i)}). \tag{2.20}$$

Another interpretation of this is that Monte Carlo methods form an approximation of the posterior density of the form

$$p(\mathbf{x} \,|\, \mathbf{y}_1, \ldots, \mathbf{y}_n) \approx \frac{1}{N} \sum_{i=1}^{N} \delta(x - x^{(i)}), \tag{2.21}$$

where $\delta(\cdot)$ is the Dirac delta function. The convergence of Monte Carlo approximation is guaranteed by the *central limit theorem (CLT)* (see, e.g., Liu, 2001) and the error term is, at least in theory, independent of the dimensionality of $\mathbf{x}$.

- Efficient methods for generating non-independent Monte Carlo samples are the *Markov chain Monte Carlo* (MCMC) methods (see, e.g., Gilks et al., 1996). In MCMC methods, a Markov chain is constructed such that it has the target distribution as its stationary distribution. By simulating the Markov chain, samples from the target distribution can be generated.

- *Importance sampling* (see, e.g., Liu, 2001) is a simple algorithm for generating *weighted* samples from the target distribution. The difference to the direct Monte Carlo sampling and to MCMC is that each of the particles contains a weight, which corrects the difference between the actual target distribution and the approximation obtained from an importance distribution $\pi(\cdot)$.

  Importance sampling estimate can be formed by drawing N samples from the *importance distribution*

$$\mathbf{x}^{(i)} \sim \pi(\mathbf{x} \,|\, \mathbf{y}_1, \ldots, \mathbf{y}_n), \qquad i = 1, \ldots, N. \tag{2.22}$$

  The *importance weights* are then computed as

$$w^{(i)} = \frac{p(\mathbf{x}^{(i)} \,|\, \mathbf{y}_1, \ldots, \mathbf{y}_n)}{\pi(\mathbf{x}^{(i)} \,|\, \mathbf{y}_1, \ldots, \mathbf{y}_n)}, \tag{2.23}$$

  and the expectation of any function $\mathbf{g}(\cdot)$ can be then approximated as

$$\mathrm{E}[\mathbf{g}(\mathbf{x}) \,|\, \mathbf{y}_1, \ldots, \mathbf{y}_n] \approx \frac{\sum_{i=1}^{N} w^{(i)} \, \mathbf{g}(\mathbf{x}^{(i)})}{\sum_{i=1}^{N} w^{(i)}}. \tag{2.24}$$

2.1.6 Notational Conventions

All probability distributions in this thesis can be represented as either probability density functions $p(\mathbf{x})$ of continuous random variables or as discrete probability distributions $P(\mathbf{x})$ of discrete random variables. Because the cardinality (i.e., discrete or continuous space) of the random variable already determines whether the probability distribution should be a density or a discrete distribution, we can simplify the notation a bit. In this thesis, the common shorthand notation (as in, e.g., Bernardo and Smith, 1994) for distributions and densities is used:

- If the random variable $\mathbf{x}$ is continuous, $p(\mathbf{x} \,|\, \mathcal{H})$ denotes the conditional probability density function of the random variable $\mathbf{x}$, given a set of hypotheses $\mathcal{H}$.
- If the random variable $\mathbf{x}$ is discrete, $p(\mathbf{x} \,|\, \mathcal{H})$ denotes the conditional discrete probability distribution of the random variable $\mathbf{x}$, given $\mathcal{H}$.
- If the random variable $\mathbf{x}$ contains both continuous and discrete components, then $p(\mathbf{x} \,|\, \mathcal{H})$ denotes the conditional hybrid probability distribution, which is density with respect to the continuous part and discrete distribution with respect to the discrete part.
- The integral notation is always used and the an integration over a distribution should be interpreted as a summation if the random variable $\mathbf{x}$ is discrete:
$$\sum_{\mathbf{x}} p(\mathbf{x} \,|\, \mathcal{H}) \triangleq \int p(\mathbf{x} \,|\, \mathcal{H}) \, \mathrm{d}\mathbf{x}. \tag{2.25}$$
This convention can also be interpreted such that when the random variable $\mathbf{x}$ is discrete and has the distribution $P(\mathbf{x})$, it has a (generalized) probability density of the form
$$p(\mathbf{x} \,|\, \mathcal{H}) = \sum_i P(\mathbf{x}_i) \, \delta(\mathbf{x} - \mathbf{x}_i), \tag{2.26}$$
where $\delta(\cdot)$ is the Dirac delta function.
- The term probability distribution is used for both probability distributions and densities.

2.2 Stochastic Differential Equations

This section briefly reviews the Itô calculus and the theory of stochastic differential equations. The presentation is quite informal and the mathematical technicalities are kept to minimum, but still rigorous enough for the purposes of theory of optimal filtering. For proofs of the theorems and details of the definitions reader is referred to Karatzas and Shreve (1991). Easier to read introductions to the subject are the lecture notes of Karatzas (1988) and the book of Øksendal (2003).

The presentation of Itô calculus is mostly based on the three references mentioned above. The definitions of Stratonovich integrals are from (Stratonovich, 1968). The notation, however, resembles more closely the notation used in more engineering related filtering theory books (e.g., Bucy and Joseph, 1968; Jazwinski, 1970) because this notation is closer to the notation used in applied Kalman filtering and stochastic control literature (e.g., Gelb, 1974; Maybeck, 1979, 1982a,b; Stengel, 1994; Grewal and Andrews, 2001; Bar-Shalom et al., 2001). In applied literature the white noise notation is often preferred to the more rigorous notation with Brownian motion.

2.2.1 Motivation

As discussed in Section 1.2, many dynamic processes in engineering, physics, finance, and other fields can be modeled as differential equations with an unknown driving function $\mathbf{w}(t)$ as follows:

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{f}(\mathbf{x}, t) + \mathbf{L}(\mathbf{x}, t)\,\mathbf{w}(t). \tag{2.27}$$

The unknown function $\mathbf{w}(t)$ would be ideally modeled as a process that is Gaussian and completely "white" in the sense that $\mathbf{w}(t)$ and $\mathbf{w}(s)$ are uncorrelated (and independent) for all $t \neq s$. However, the problem is that this kind of process cannot exists in any mathematically or physically meaningful sense (Øksendal, 2003).

The solution to this existence problem is that actually the white process does not need to exist as long as its integral exists. Integrating the Equation (2.27) once with respect to time gives the stochastic integral equation

$$\mathbf{x}(t) - \mathbf{x}(s) = \int_s^t \mathbf{f}(\mathbf{x}, t)\,\mathrm{d}t + \int_s^t \mathbf{L}(\mathbf{x}, t)\,\mathbf{w}(t)\,\mathrm{d}t \tag{2.28}$$

The first integral on the right hand side does not cause any problems, but the second integral is problematic because of the appearance of white noise process. Fortunately, this integral can be defined to be a integral with respect to the stochastic "measure" $\beta(t)$, which has independent Gaussian increments:

$$\int_s^t \mathbf{L}(\mathbf{x}, t)\,\mathbf{w}(t)\,\mathrm{d}t \triangleq \int_s^t \mathbf{L}(\mathbf{x}, t)\,\mathrm{d}\boldsymbol{\beta}(t). \tag{2.29}$$

The process $\boldsymbol{\beta}(t)$ is the Brownian motion to be defined in the next section. This kind of integral is called a stochastic or Itô integral. *White noise* is then, at least in formal sense, the time derivative of the Brownian motion $\mathbf{w}(t) = \mathrm{d}\boldsymbol{\beta}(t)/\mathrm{d}t$.

Because by a stochastic differential equation it is actually meant the corresponding stochastic integral equation, this point is emphasized by writing stochastic differential equations in form

$$\mathrm{d}\mathbf{x}(t) = \mathbf{f}(\mathbf{x}, t)\,\mathrm{d}t + \mathbf{L}(\mathbf{x}, t)\,\mathrm{d}\boldsymbol{\beta}(t). \tag{2.30}$$

where the usage of the problematic white noise process is avoided. The next sections define what is actually meant by the integral with respect to Brownian motion.

2.2.2 Stochastic Processes, Brownian Motion and Martingales

Definition 2.2 (Stochastic process)**.** *An indexed collection of random variables*

$$\mathcal{X}(\omega) = \{\mathbf{x}(t;\omega), 0 \le t < \infty\}, \tag{2.31}$$

is called a stochastic process.

- Each $\omega \mapsto \mathbf{x}(t;\omega)$ is a measurable function defined on a probability space $(\Omega, \mathscr{A}, P)$.
- For each $\omega \in \Omega$ the function $t \mapsto \mathbf{x}(t;\omega)$ is called the *sample path* (or realization or trajectory) of the process.

Definition 2.3 (Filtration)**.** *The increasing family of σ-algebras $\mathscr{X}_t \subset \mathscr{A}$ on Ω such that*

$$0 \le s < t \Rightarrow \mathscr{X}_s \subset \mathscr{X}_t, \tag{2.32}$$

is called a filtration. *The stochastic process $\mathbf{x}(t;\omega)$ is said to be* adapted *to the filtration $\mathscr{X}_t$ if for each $t \ge 0$ the function $\omega \mapsto \mathbf{x}(t;\omega)$ is $\mathscr{X}_t$-measurable. The* natural filtration *of a stochastic process is the smallest filtration such that the process is adapted to it.*

The natural filtration $\mathscr{X}_t$ of stochastic process $\mathbf{x}(t;\omega)$, that is,

$$\mathscr{X}_t = \sigma\left(\mathbf{x}(s), 0 \le s < t\right), \tag{2.33}$$

can be thought of as the history of the stochastic process up to the time t. The filtration contains all the information that can be known about the process at the time t.

Definition 2.4 (Markov process)**.** *A stochastic process* $\mathbf{x}(t)$ *is a* Markov process *if its future is independent of its past given the present:*

$$p(\mathbf{x}(s) \mid \mathscr{X}_t) = p(\mathbf{x}(s) \mid \mathbf{x}(t)), \text{ for all } s \geq t. \tag{2.34}$$

Definition 2.5 (Martingale)**.** *An* $\mathscr{X}_t$*-adapted stochastic process* $\mathbf{x}(t)$ *with bounded expectation* $\mathrm{E}[\mathbf{x}(t)] < \infty$ *is called a* martingale *with respect to the filtration* $\mathscr{X}_t$ *if*

$$\mathrm{E}[\mathbf{x}(s) \mid \mathscr{X}_t] = \mathbf{x}(t), \text{ for all } s \geq t. \tag{2.35}$$

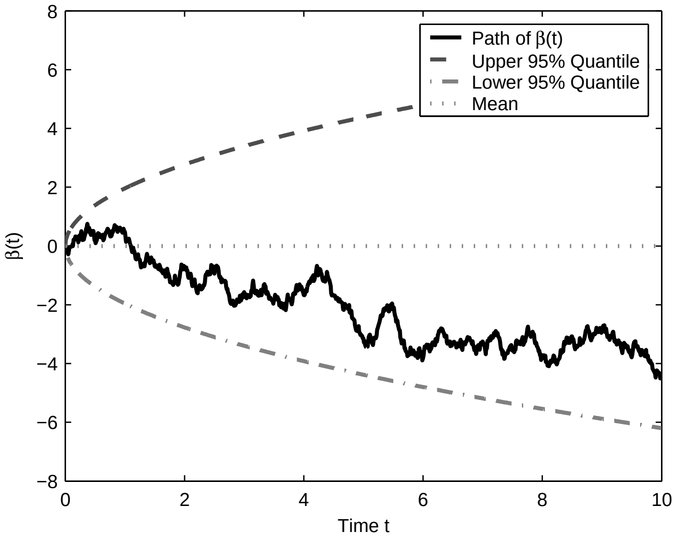

Figure 2.1: Single trajectory (realization) of Brownian motion.

Definition 2.6 (Standard Brownian motion)**.** *A process* $\beta(t)$ *is called a* standard Brownian motion[1] *if it has the following properties:*

1. $\beta(0) = 0$.
2. $\beta(t_1), \beta(t_2) - \beta(t_1), \ldots, \beta(t_k) - \beta(t_{k-1})$ *are independent for all* $t_1 < t_2 < \ldots < t_{k-1} < t_k < \infty$.
3. $\beta(t) - \beta(s) \sim \mathrm{N}(0, t-s)$ *for every* $0 < s < t < \infty$.
4. *The sample path* $t \mapsto \beta(t; \omega)$ *is continuous for all* $\omega \in \Omega$.

[1] Standard Brownian motion is also called Wiener process

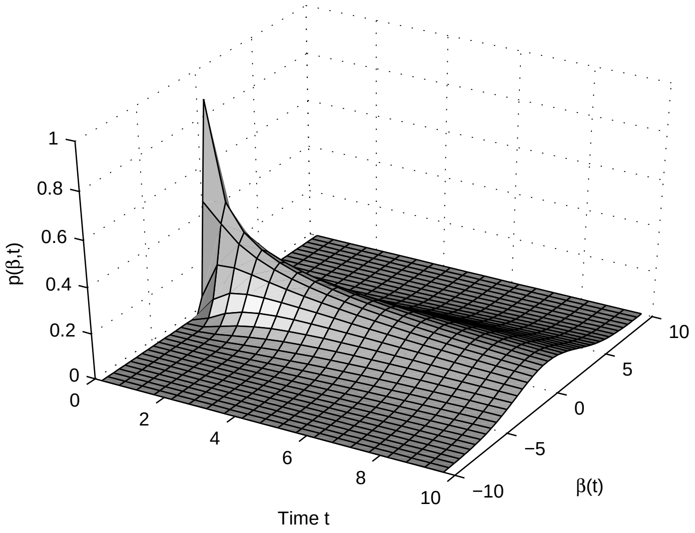

Figure 2.2: Probability density of Brownian motion.

An n-dimensional vector process $\boldsymbol{\beta}(t) = (\beta_1(t) \;\cdots\; \beta_n(t))^T$ where each scalar process $\beta_i(t)$ is an independent standard Brownian motion is called an n-dimensional standard Brownian motion.

Definition 2.7 (Brownian motion)**.** *A (scalar)* Brownian motion with diffusion coefficient $q(t)$ *can be defined as the process*

$$\beta(t) = \sqrt{q(t)}\,\beta_s(t), \tag{2.36}$$

where $\beta_s(t)$ is a standard Brownian motion. An n-dimensional vector process $\boldsymbol{\beta}(t) = (\beta_1(t) \;\cdots\; \beta_n(t))^T$ where $\beta_i(t)$ are independent Brownian motions with diffusion coefficients $q_i(t)$ is called n-dimensional Brownian motion *with diffusion matrix* $\mathbf{Q}_c(t) = \operatorname{diag}(q_1(t), \ldots, q_n(t))$.

2.2.3 Stochastic Integral

In this section we shall define the Itô integral, that is, the stochastic integral of a function (or stochastic process) f with respect to Brownian motion $\beta(t)$

$$\mathcal{I}[f] = \int_S^T f(t, \omega)\, \mathrm{d}\beta. \tag{2.37}$$

We shall also define the multidimensional generalization of the integral. The symmetrized stochastic integral, which is called the Stratonovich integral is also discussed.

Definition 2.8 (Simple process). *Stochastic process $\phi_n(t,\omega) : [S,T] \times \Omega \to \mathbb{R}$ is called* simple *if there exists partition $S = t_0 < t_1 < \ldots < t_n < t_{n+1} = T$ such that $\phi_n(s,\omega) = \theta_j(\omega), t_j < s \le t_{j+1}$ where $\theta_j(\omega)$ is a random variable. For technical reasons we shall also require that each $\theta_j(\omega)$ is measurable with respect to a filtration $\mathscr{H}_t$ such the Brownian motion $\beta(t)$ is martingale with respect to the filtration.*

Thus, simple process is a piecewise constant stochastic process. For simple processes the Itô integral can be defined as follows:

$$\begin{aligned}\mathfrak{I}[\phi_n] &= \int_S^T \phi_n(t,\omega)\,\mathrm{d}\beta \\ &= \sum_{j=0}^{n} \theta_j(\omega)\,(\beta(t_{j+1}) - \beta(t_j)).\end{aligned} \tag{2.38}$$

Itô integral of a more general stochastic process can be now defined as limit of integrals of simple processes:

Definition 2.9 (Itô integral). *Let $(\Omega, \mathscr{A}, P)$ be a probability space, $\beta(t)$ a Brownian motion with natural filtration $\mathscr{F}_t \subset \mathscr{A}$ and $f(t,\omega) : [0,\infty) \times \Omega \mapsto \mathbb{R}$ a stochastic process with the following properties:*

1. *$(t,\omega) \mapsto f(t,\omega)$ is $\mathscr{B}[0,\infty) \times \mathscr{A}$-measurable.*
2. *There exists filtration $\mathscr{H}_t$ such that $\beta(t)$ is martingale with respect to $\mathscr{H}_t$ and $f(t,\omega)$ is $\mathscr{H}_t$-adapted.*
3. *$\mathrm{E}[\int_S^T f(t,\omega)^2\,\mathrm{d}t] < \infty$.*

Then the Itô integral *of $f(t,\omega)$ with respect to the Brownian motion $\beta(t)$ can be defined as*

$$\int_S^T f(t,\omega)\,\mathrm{d}\beta(t;\omega) = \lim_{n\to\infty} \int_S^T \phi_n(t,\omega)\,\mathrm{d}\beta(t;\omega), \tag{2.39}$$

where $\{\phi_n\}$ is a sequence of simple processes such that

$$\lim_{n\to\infty} \mathrm{E}\left[\int_S^T (f(t,\omega) - \phi_n(t,\omega))\,\mathrm{d}t\right] = 0. \tag{2.40}$$

Note that the Itô integral is always a martingale.

Definition 2.10 (Multi-dimensional Itô integral). *: Let $\boldsymbol{\beta}(t) = (\beta_1(t), \ldots, \beta_n(t))$ be a n-dimensional Brownian motion and $\mathbf{L}(t;\omega)$ a matrix valued process where each component L_{ij} satisfies 1-3 in the definition 2.9. The Itô integral of $\mathbf{L}(t;\omega)$*

with respect to the n-dimensional Brownian motion can be defined to be a vector, where the i'th component is given as

$$\sum_{j=1}^{n} \int_S^T L_{ij}(t;\omega)\,\mathrm{d}\beta_j. \tag{2.41}$$

Itô integrals can also be defined with respect to more general martingales than Brownian motion (Karatzas and Shreve, 1991; Applebaum, 2004). This kind of processes are for example general Lévy processes of which the Poisson process (and also the Brownian motion) is a special case. However, in this thesis only stochastic integrals and stochastic differential equations defined in terms of Brownian motions are considered.

The symmetrized stochastic integral or the *Stratonovich integral* (Stratonovich, 1968) can be defined as a limit of the symmetrized integrals of simple functions defined as follows:

$$\begin{aligned} \mathcal{S}[\phi_n] &= \int_S^T \phi_n(t,\omega) \circ \mathrm{d}\beta \\ &= \sum_{j=0}^{n} \left(\frac{\theta_j(\omega) + \theta_{j+1}(\omega)}{2} \right) (\beta(t_{j+1}) - \beta(t_j)). \end{aligned} \tag{2.42}$$

The Stratonovich integral can be defined in analogous manner as the Itô integral, but with a bit less general conditions for the process $f(t,\omega)$.

The disadvantage of the Stratonovich integral over Itô integral is that the Stratonovich integral is not a martingale. However, the advantage and also the reason for its development is that the normal rules of calculus apply, when the Stratonovich interpretation of stochastic integrals is used.

2.2.4 Stochastic Differential Equations

Definition 2.11 (Stochastic differential equation)**.** Stochastic differential equation *(SDE) is an equation of the form*

$$\mathrm{d}\mathbf{x}(t) = \mathbf{f}(\mathbf{x},t)\,\mathrm{d}t + \mathbf{L}(\mathbf{x},t)\,\mathrm{d}\boldsymbol{\beta}(t), \tag{2.43}$$

where $\mathbf{f} : \mathbb{R}^n \times [0,\infty) \mapsto \mathbb{R}^n$ *is the* drift function , $\mathbf{L} : \mathbb{R}^n \times [0,\infty) \mapsto \mathbb{R}^{n\times d}$ *is the* dispersion matrix *and* $\boldsymbol{\beta}(t)$ *is a* d*-dimensional Brownian motion with diffusion matrix* $\mathbf{Q}_c(t)$*. The matrix* $\mathbf{L}(\mathbf{x},t)\,\mathbf{Q}_c(t)\,\mathbf{L}^T(\mathbf{x},t)$ *is then the diffusion matrix of the stochastic differential equation. The stochastic differential equation* (2.43) *is actually a short hand notation to the stochastic integral equation*

$$\mathbf{x}(t) - \mathbf{x}(s) = \int_s^t \mathbf{f}(\mathbf{x},t)\,\mathrm{d}t + \int_s^t \mathbf{L}(\mathbf{x},t)\,\mathrm{d}\boldsymbol{\beta}(t), \tag{2.44}$$

where the last integral is an Itô stochastic integral. The stochastic process solution $\mathbf{x}(t)$ *to the stochastic differential equation is called* Itô process .

The *Stratonovich stochastic differential equations* (Stratonovich, 1968; Øksendal, 2003) are similar to Itô differential equations, but instead of Itô integrals they involve stochastic integrals in the Stratonovich sense. A Stratonovich stochastic differential equation can always be converted into an equivalent Itô equation by using simple transformation formulas (Stratonovich, 1968; Øksendal, 2003). If the dispersion term is independent of the state $\mathbf{L}(\mathbf{x},t) = \mathbf{L}(t)$ then the Itô and Stratonovich interpretations of the stochastic differential equation are the same.

To distinguish between Itô and Stratonovich stochastic differential equations, the Stratonovich integral is denoted by a small circle before the Brownian differential as follows:

$$\mathrm{d}\mathbf{x}(t) = \mathbf{f}(\mathbf{x},t)\,\mathrm{d}t + \mathbf{L}(\mathbf{x},t)\circ\mathrm{d}\boldsymbol{\beta}(t). \tag{2.45}$$

The white noise interpretation of SDEs naturally leads to stochastic differential equations in Stratonovich sense. This is because the discrete-time approximations of white noise driven differential equations converge to stochastic differential equations in Stratonovich sense, not in Itô sense. For this reason higher order numerical integration schemes also approximate the corresponding Stratonovich equation when applied to stochastic differential equations.

A solution to a stochastic differential equation is called *strong* if for given Brownian motion $\boldsymbol{\beta}(t)$ with filtration $\mathscr{F}_t$ it is possible to construct a solution $\mathbf{x}(t)$, which is $\mathscr{F}_t$-adapted. Uniqueness of a strong solution means that the paths of the process are unique for given Brownian motion and for this reason strong uniqueness is also called path-wise uniqueness.

A solution is called *weak* if it is possible to construct some Brownian motion $\hat{\boldsymbol{\beta}}(t)$ and a stochastic process $\hat{\mathbf{x}}(t)$ such that the pair is a solution to the stochastic differential equation. Weak uniqueness means that the probability law of the solution is unique, that is, there cannot be two solutions with different finite-dimensional distributions.

The required conditions for drift function $\mathbf{f}$ and dispersion matrix $\mathbf{L}$, which guarantee existences of strong and weak solutions can be found in the books of Karatzas and Shreve (1991) and Øksendal (2003).

The most important tool for computing strong solutions to stochastic differential equations is the Itô formula, which can be interpreted as counterpart of the chain rule in ordinary calculus:

Theorem 2.1 (Itô formula)**.** *Assume that the process $\mathbf{x}(t)$ is generated by the stochastic differential equation*

$$\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x},t)\,\mathrm{d}t + \mathbf{L}(\mathbf{x},t)\,\mathrm{d}\boldsymbol{\beta}. \tag{2.46}$$

where $\boldsymbol{\beta}(t)$ is a Brownian motion with diffusion matrix $\mathbf{Q}_c(t)$. Let $\mathbf{g}$ be a twice differentiable function. Then the components of the stochastic process $\mathbf{y}(t) =$

$\mathbf{g}(\mathbf{x}(t), t)$ *satisfy the stochastic differential equations*

$$\mathrm{d}y_k = \frac{\partial g_k}{\partial t}\,\mathrm{d}t + \sum_i \frac{\partial g_k}{\partial x_i}\,\mathrm{d}x_i + \frac{1}{2}\sum_{ij} \frac{\partial^2 g_k}{\partial x_i\,\partial x_j}\mathrm{d}x_i\,\mathrm{d}x_j, \tag{2.47}$$

where the terms $\mathrm{d}x_i\,\mathrm{d}x_j$ *are computed according to the rules*

$$\begin{aligned} \mathrm{d}t\,\mathrm{d}\boldsymbol{\beta} &= \mathbf{0} \\ \mathrm{d}\boldsymbol{\beta}\,\mathrm{d}t &= \mathbf{0} \\ \mathrm{d}\boldsymbol{\beta}\,\mathrm{d}\boldsymbol{\beta}^T &= \mathbf{Q}_c(t)\,\mathrm{d}t. \end{aligned} \tag{2.48}$$

Remark 2.1 (Stratonovich formula)**.** *If the Equation* (2.46) *was a stochastic differential equation in Stratonovich sense*

$$\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\,\mathrm{d}t + \mathbf{L}(\mathbf{x}, t) \circ \mathrm{d}\boldsymbol{\beta}, \tag{2.49}$$

then the differential would be

$$\mathrm{d}y_k = \frac{\partial g_k}{\partial t}\,\mathrm{d}t + \sum_i \frac{\partial g_k}{\partial x_i} \circ \mathrm{d}x_i. \tag{2.50}$$

That is, the familiar result from calculus.

In Bayesian inference all information about the unknown quantities is assumed to be contained in the probability distribution of the unknown quantities. For this reason, when doing Bayesian inference on stochastic differential equations weak solutions to stochastic differential equations are often enough, because we are only interested in the probability laws, not the actual paths of the processes.

The probability distribution, that is, the law of any weak solution to a stochastic differential equation can be computed by the Kolmogorov forward equation:

Theorem 2.2 (Kolmogorov forward equation)**.** *The probability density of the stochastic process* $\mathbf{x}(t)$ *which is generated by the differential equation*

$$\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\,\mathrm{d}t + \mathbf{L}(\mathbf{x}, t)\,\mathrm{d}\boldsymbol{\beta}. \tag{2.51}$$

satisfies the Kolmogorov forward partial differential equation*, also called the* Fokker-Planck equation

$$\frac{\partial p}{\partial t} = -\sum_i \frac{\partial}{\partial x_i}\left(f_i(\mathbf{x}, t)\,p\right) + \frac{1}{2}\sum_{ij} \frac{\partial^2}{\partial x_i \partial x_j}\left([\mathbf{L}(\mathbf{x}, t)\,\mathbf{Q}(t)\,\mathbf{L}^T(\mathbf{x}, t)]_{ij}\,p\right), \tag{2.52}$$

where the probability density $p(\mathbf{x}(t)) = p(\mathbf{x}, t)$ *is interpreted as function of* $\mathbf{x}$ *and* t*. The equation can also be written in the operator form*

$$\frac{\partial p}{\partial t} = \mathcal{A}_t^*[p], \tag{2.53}$$

where the operator $\mathcal{A}_t^*$ is

$$\mathcal{A}_t^*[\cdot] = -\sum_i \frac{\partial}{\partial x_i} \left(f_i(\mathbf{x}, t) \, (\cdot) \right) + \frac{1}{2} \sum_{ij} \frac{\partial^2}{\partial x_i \partial x_j} \left([\mathbf{L}(\mathbf{x}, t) \, \mathbf{Q}(t) \, \mathbf{L}^T(\mathbf{x}, t)]_{ij} \, (\cdot) \right), \tag{2.54}$$

which is the formal adjoint *of the characteristic operator $\mathcal{A}_t$ of the diffusion process.*

Example 2.1 (Diffusion equation)**.** *An interesting connection to statistical physics is that a plain standard Brownian motion can be defined by the SDE*

$$\mathrm{d}x = \mathrm{d}\beta. \tag{2.55}$$

Now the Kolmogorov forward equation reduces to

$$\frac{\partial p}{\partial t} = \frac{1}{2} \frac{\partial^2 p}{\partial x^2}, \tag{2.56}$$

which can be recognized as the diffusion equation of statistical physics. This connection between the Brownian motion and the diffusion equation was already known by Einstein (1905).

2.2.5 Girsanov Theorem

Theorem 2.3 (Girsanov)**.** *Assume that $\{\boldsymbol{\theta}(t) \in \mathbb{R}^n \; : \; 0 \leq t \leq T\}$ is a $\mathscr{A}$-measurable process, which is adapted to the natural filtration $\mathscr{F}_t \subset \mathscr{A}$ of n-dimensional standard Brownian motion $\{\boldsymbol{\beta}(t) \in \mathbb{R}^n \; : \; 0 \leq t \leq T\}$ with respect to measure P. If*

$$\mathrm{E}\left[\exp\left(\int_0^t ||\boldsymbol{\theta}(t)||^2 \, \mathrm{d}t \right) \right] < \infty, \tag{2.57}$$

then

$$Z(t) = \exp\left(\int_0^t \boldsymbol{\theta}^T(t) \, \mathrm{d}\boldsymbol{\beta}(t) - \frac{1}{2} \int_0^t ||\boldsymbol{\theta}(t)||^2 \, \mathrm{d}t \right), \tag{2.58}$$

satisfies the equation

$$Z(t) = 1 + \int_0^t Z(t) \, \boldsymbol{\theta}^T(t) \, \mathrm{d}\beta(t), \tag{2.59}$$

and is a martingale. Then under the measure $\tilde{P}(\mathrm{d}\omega) = Z(t; \omega) \, P(\mathrm{d}w)$ the process

$$\tilde{\boldsymbol{\beta}}(t) = \boldsymbol{\beta}(t) - \int_0^t \boldsymbol{\theta}(t) \, \mathrm{d}t, \tag{2.60}$$

is n-dimensional standard Brownian motion. The random variable $Z(t; \omega)$ is the likelihood ratio *between laws $\tilde{P}$ and P*

$$\left. \frac{\mathrm{d}\tilde{P}}{\mathrm{d}P}(\omega) \right|_{\mathscr{F}_t} = Z(t; \omega). \tag{2.61}$$

Proof. See (Karatzas and Shreve, 1991; Øksendal, 2003). □

The Girsanov theorem can readily be applied to finding weak solutions and for removing drifts from stochastic differential equations of the form

$$\begin{aligned} \mathrm{d}\mathbf{x} &= \mathbf{f}(\mathbf{x},t)\,\mathrm{d}t + \mathbf{L}\,\mathrm{d}\boldsymbol{\beta} \\ \mathbf{x}(0) &= \mathbf{x}_0, \end{aligned} \tag{2.62}$$

where $\mathbf{f}(\mathbf{x},t)$ is bounded and measurable, $\mathbf{L}$ is invertible matrix and $\boldsymbol{\beta}(t)$ is a Brownian motion with respect to measure P.

Theorem 2.4 (Weak solution of SDE). *Assume that the process $\mathbf{x}(t)$ is generated by the stochastic differential equation* (2.62). *If we now define*

$$\begin{aligned} Z(t;\omega) = \exp\Bigg(&\int_0^t \left[\mathbf{L}^{-1}\,\mathbf{f}(\mathbf{x}_0 + \mathbf{L}\,\boldsymbol{\beta}(t;\omega),t)\right]^T\,\mathrm{d}\boldsymbol{\beta}(t;\omega) \\ &- \frac{1}{2}\int_0^t ||\mathbf{L}^{-1}\,\mathbf{f}(\mathbf{x}_0 + \mathbf{L}\,\boldsymbol{\beta}(t;\omega),t)||^2\,\mathrm{d}t\Bigg), \end{aligned} \tag{2.63}$$

then the expectation of any function (or functional) $\mathbf{h}(\cdot)$ can be expressed as

$$\mathrm{E}[\mathbf{h}(\mathbf{x}(t))] = \mathrm{E}[Z(t;\omega)\,\mathbf{h}(\mathbf{x}_0 + \mathbf{L}\,\boldsymbol{\beta}(t;\omega))]. \tag{2.64}$$

and thus $Z(t;\omega)$ is the likelihood ratio between processes $\mathbf{x}(t)$ and $\mathbf{x}_0 + \mathbf{L}\,\boldsymbol{\beta}(t;\omega)$.

Proof. If we define

$$\tilde{\mathbf{x}}(t) = \mathbf{x}_0 + \mathbf{L}\,\boldsymbol{\beta}(t;\omega) \tag{2.65}$$

$$\boldsymbol{\theta}(t) = \mathbf{L}^{-1}\,\mathbf{f}(\mathbf{x}_0 + \mathbf{L}\,\boldsymbol{\beta}(t),t) \tag{2.66}$$

$$\begin{aligned} \tilde{\boldsymbol{\beta}}(t) &= \boldsymbol{\beta}(t) - \int_0^t \boldsymbol{\theta}(t)\,\mathrm{d}t \\ &= \mathbf{L}^{-1}\,(\tilde{\mathbf{x}}(t) - \mathbf{x}_0) - \int_0^t \mathbf{L}^{-1}\,\mathbf{f}(\tilde{\mathbf{x}}(t),t)\,\mathrm{d}t. \end{aligned} \tag{2.67}$$

then by rearranging the last equation we get that the processes $\tilde{\mathbf{x}}(t), \tilde{\boldsymbol{\beta}}(t)$ satisfy

$$\mathrm{d}\tilde{\mathbf{x}} = \mathbf{f}(\tilde{\mathbf{x}},t)\,\mathrm{d}t + \mathbf{L}\,\mathrm{d}\tilde{\boldsymbol{\beta}}(t) \tag{2.68}$$

$$\tilde{\mathbf{x}}(0) = \mathbf{x}_0. \tag{2.69}$$

By the Girsanov theorem, under the measure $\tilde{P}(\mathrm{d}\omega) = Z(t;\omega)\,P(\mathrm{d}\omega)$ the process $\tilde{\boldsymbol{\beta}}(t)$ is a Brownian motion and thus the pair $(\tilde{\mathbf{x}}(t), \tilde{\boldsymbol{\beta}}(t))$ is a weak solution to the SDE. For any function $\mathbf{h}(\cdot)$ we now have

$$\tilde{\mathrm{E}}[\mathbf{h}(\tilde{\mathbf{x}}(t))] = \mathrm{E}[Z(t;\omega)\,\mathbf{h}(\mathbf{x}_0 + \mathbf{L}\,\boldsymbol{\beta}(t;\omega))]. \tag{2.70}$$

But by definition of a weak solution we should have $\tilde{\mathrm{E}}[\mathbf{h}(\tilde{\mathbf{x}}(t))] = \mathrm{E}[\mathbf{h}(\mathbf{x}(t))]$ and the result follows. □

Theorem 2.5 (Removal of drift)**.** *Assume that the process $\mathbf{x}(t)$ is generated by the stochastic differential equation* (2.62)*. If we define*

$$Z(t;\omega) = \exp\left(- \int_0^t \left[\mathbf{L}^{-1}\mathbf{f}(\mathbf{x}(t),t)\right]^T \mathrm{d}\boldsymbol{\beta}(t) - \frac{1}{2}\int_0^t ||\mathbf{L}^{-1}\mathbf{f}(\mathbf{x}(t),t)||^2 \mathrm{d}t \right), \tag{2.71}$$

Then under the measure $\tilde{P}(\mathrm{d}\omega) = Z(t;\omega)\,P(\mathrm{d}\omega)$ the process $\mathbf{x}(t) - \mathbf{x}_0$ is a Brownian motion with diffusion matrix $\mathbf{L}\,\mathbf{L}^T$ and thus the law of $\mathbf{x}(t)$ is the same as the law of $\mathbf{x}_0 + \mathbf{L}\,\boldsymbol{\beta}(t)$.

Proof. If we define

$$\boldsymbol{\theta}(t) = -\mathbf{L}^{-1}\mathbf{f}(\mathbf{x}(t),t) \tag{2.72}$$

$$\tilde{P}(\mathrm{d}\omega) = Z(t;\omega)\,P(\mathrm{d}\omega), \tag{2.73}$$

then by Theorem 2.3 we get that under the measure $\tilde{P}$

$$\begin{aligned}
\tilde{\boldsymbol{\beta}}(t) &= \boldsymbol{\beta}(t) - \int_0^t \boldsymbol{\theta}(t)\,\mathrm{d}t \\
&= \mathbf{L}^{-1}\left(\mathbf{x}(t) - \mathbf{x}_0\right) - \int_0^t \mathbf{L}^{-1}\mathbf{f}(\mathbf{x}(t),t)\,\mathrm{d}t + \int_0^t \mathbf{L}^{-1}\mathbf{f}(\mathbf{x}(t),t)\,\mathrm{d}t \\
&= \mathbf{L}^{-1}(\mathbf{x}(t) - \mathbf{x}_0)
\end{aligned}$$

is a standard Brownian motion and thus $\mathbf{x}(t) - \mathbf{x}_0$ is a Brownian motion with diffusion matrix $\mathbf{L}\,\mathbf{L}^T$. □

Example 2.2 (Solution of Beneš SDE)**.** *Consider the Itô process*

$$\mathrm{d}x = \tanh(x)\,\mathrm{d}t + \mathrm{d}\beta(t) \tag{2.74}$$

$$x(0) = x_0, \tag{2.75}$$

where $\beta(t)$ is a standard Brownian motion. The likelihood ratio is now

$$Z(t) = \exp\left(\int_0^t \tanh(x_0 + \beta(t))\,\mathrm{d}\beta(t) - \frac{1}{2}\int_0^t \tanh^2(x_0 + \beta(t))\,\mathrm{d}t\right). \tag{2.76}$$

The first integral can be evaluated by applying the Itô formula:

$$\begin{aligned}
&\mathrm{d}(\log(\cosh(x_0 + \beta(t)))) \\
&\quad = \tanh(x_0 + \beta(t))\,\mathrm{d}\beta(t) + \frac{1}{2}(1 - \tanh^2(x_0 + \beta(t)))\,\mathrm{d}t,
\end{aligned} \tag{2.77}$$

that is

$$\begin{aligned}\int_0^t \tanh(x_0 + \beta(t))\,\mathrm{d}\beta(t) - \frac{1}{2}\int_0^t \tanh^2(x_0 + \beta(t)))\,\mathrm{d}t \\ = \log(\cosh(x_0 + \beta(t))) - \log(\cosh(x_0)) - \int_0^t \frac{1}{2}\,\mathrm{d}t,\end{aligned} \tag{2.78}$$

which gives

$$\begin{aligned}Z(t) &= \exp\left(\log(\cosh(x_0 + \beta(t))) - \log(\cosh(x_0)) - \int_0^t \frac{1}{2}\,\mathrm{d}t\right) \\ &= \frac{\cosh(x_0 + \beta(t))}{\cosh(x_0)}\,\exp\left(-\frac{1}{2}t\right).\end{aligned} \tag{2.79}$$

Because this term depends only on the state of Brownian motion at time t, the probability density can be computed explicitly. The probability density of $\tilde{x}(t) = x_0 + \beta(t)$ is

$$\mathrm{N}(\tilde{x}(t)\,|\,0, t) = \frac{1}{\sqrt{2\pi t}}\exp\left(-\frac{1}{2t}(\tilde{x}(t) - x_0)^2\right), \tag{2.80}$$

and thus the probability density of $x(t)$ is

$$p(x(t)) = \frac{1}{\sqrt{2\pi t}}\frac{\cosh(x(t))}{\cosh(x_0)}\,\exp\left(-\frac{1}{2}t\right)\,\exp\left(-\frac{1}{2t}(x(t) - x_0)^2\right). \tag{2.81}$$

Because this is true for any initial condition, it implies that the transition density is given as

$$\begin{aligned}p(x(t_k)\,|\,x(t_{k-1})) = \frac{1}{\sqrt{2\pi\Delta t_k}}\frac{\cosh(x(t_k))}{\cosh(x(t_{k-1}))}\,\exp\left(-\frac{1}{2}\Delta t_k\right) \\ \times \exp\left(-\frac{1}{2\Delta t_k}(x(t_k) - x(t_{k-1}))^2\right),\end{aligned} \tag{2.82}$$

where $\Delta t_k = t_k - t_{k-1}$.

2.2.6 Solutions of Linear Equations

Theorem 2.6 (Solution of linear SDE)**.** Linear stochastic differential equation *of the form*

$$\mathrm{d}\mathbf{x} = \mathbf{F}(t)\,\mathbf{x}(t)\,\mathrm{d}t + \mathbf{u}(t) + \mathbf{L}(t)\,\boldsymbol{\beta}(t), \tag{2.83}$$

where

- *the initial conditions are $\mathbf{x}(0) \sim \mathrm{N}(\mathbf{m}(0), \mathbf{P}(0))$,*

- $\mathbf{F}(t)$ *and* $\mathbf{L}(t)$ *are matrix valued functions,*
- $\mathbf{u}(t)$ *is a known deterministic (non-random) function,*
- *and* $\boldsymbol{\beta}(t)$ *is a Brownian motion with diffusion matrix* $\mathbf{Q}_c(t)$*,*

can be solved exactly using the ordinary differential equations

$$\frac{\mathrm{d}\mathbf{m}(t)}{\mathrm{d}t} = \mathbf{F}(t)\,\mathbf{m}(t) + \mathbf{u}(t) \tag{2.84}$$

$$\frac{\mathrm{d}\mathbf{P}(t)}{\mathrm{d}t} = \mathbf{F}(t)\,\mathbf{P}(t) + \mathbf{P}(t)\,\mathbf{F}^T(t) + \mathbf{L}(t)\,\mathbf{Q}_c(t)\,\mathbf{L}^T(t). \tag{2.85}$$

The solution is a Gaussian process with mean $\mathbf{m}(t)$ *and covariance* $\mathbf{P}(t)$*:*

$$p(\mathbf{x}(t)) = \mathrm{N}(\mathbf{x}(t)\,|\,\mathbf{m}(t), \mathbf{P}(t)). \tag{2.86}$$

Theorem 2.7 (Solution of LTI SDE)**.** *The solution of* linear time-invariant stochastic differential equation *(LTI SDE) of the form*

$$\mathrm{d}\mathbf{x} = \mathbf{F}\,\mathbf{x}(t)\,\mathrm{d}t + \mathbf{L}\,\boldsymbol{\beta}(t), \tag{2.87}$$

where

- *the initial conditions are* $\mathbf{x}(0) \sim \mathrm{N}(\mathbf{m}(0), \mathbf{P}(0))$*,*
- $\mathbf{F}$ *and* $\mathbf{L}$ *are constant matrices,*
- *and* $\boldsymbol{\beta}(t)$ *is a Brownian motion with constant diffusion matrix* $\mathbf{Q}_c$*,*

is a Gaussian process with the following mean $\mathbf{m}(t)$ *and covariance* $\mathbf{P}(t)$*:*

$$\mathbf{m}(t) = \exp(\mathbf{F}\,t)\,\mathbf{m}(0) \tag{2.88}$$

$$\begin{aligned}\mathbf{P}(t) &= \exp(\mathbf{F}\,t)\,\mathbf{P}(0)\,\exp(\mathbf{F}\,t)^T \\ &\quad + \int_0^t \exp(\mathbf{F}\,(t-\tau))\,\mathbf{L}\,\mathbf{Q}_c\mathbf{L}^T\,\exp(\mathbf{F}\,(t-\tau))^T\,\mathrm{d}\tau,\end{aligned} \tag{2.89}$$

where $\exp(\cdot)$ *is the matrix exponential function.*

Remark 2.2 (Matrix fraction decomposition)**.** *The covariance of linear time invariant*[2] *stochastic differential equation* (2.89) *can be solved by using matrix fractions (see, e.g., Stengel, 1994; Grewal and Andrews, 2001). If we define matrices*

[2] Actually the method applies to the time varying case also, but in that case the benefit is smaller because the solution cannot be expressed in terms of the matrix exponential function.

$\mathbf{C}$ *and* $\mathbf{D}$ *such that* $\mathbf{P} = \mathbf{C}\,\mathbf{D}^{-1}$*, it is easy to show that* $\mathbf{P}$ *solves the matrix Riccati differential equation*

$$\frac{\mathrm{d}\mathbf{P}(t)}{\mathrm{d}t} = \mathbf{F}\,\mathbf{P}(t) + \mathbf{P}(t)\,\mathbf{F}^T + \mathbf{L}\,\mathbf{Q}_c\,\mathbf{L}^T. \tag{2.90}$$

if matrices $\mathbf{C}$ *and* $\mathbf{D}$ *solve the differential equation*

$$\begin{pmatrix} \mathrm{d}\mathbf{C}(t)/\mathrm{d}t \\ \mathrm{d}\mathbf{D}(t)/\mathrm{d}t \end{pmatrix} = \begin{pmatrix} \mathbf{F} & \mathbf{L}\,\mathbf{Q}_c\,\mathbf{L}^T \\ \mathbf{0} & -\mathbf{F}^T \end{pmatrix} \begin{pmatrix} \mathbf{C}(t) \\ \mathbf{D}(t) \end{pmatrix}, \tag{2.91}$$

and $\mathbf{P}(0) = \mathbf{C}(0)\,\mathbf{D}(0)^{-1}$*. We can select, for example,*

$$\mathbf{C}(0) = \mathbf{P}(0) \tag{2.92}$$

$$\mathbf{D}(0) = \mathbf{I}. \tag{2.93}$$

Because the differential equation (2.91) *is linear and time invariant, it can be solved using the matrix exponential function:*

$$\begin{pmatrix} \mathbf{C}(t) \\ \mathbf{D}(t) \end{pmatrix} = \exp\left\{ \begin{pmatrix} \mathbf{F} & \mathbf{L}\,\mathbf{Q}_c\,\mathbf{L}^T \\ \mathbf{0} & -\mathbf{F}^T \end{pmatrix} t \right\} \begin{pmatrix} \mathbf{C}(0) \\ \mathbf{D}(0) \end{pmatrix}. \tag{2.94}$$

The final solution is then given as $\mathbf{P}(t) = \mathbf{C}(t)\,\mathbf{D}(t)^{-1}$.

Theorem 2.8 (Discretization of linear SDE)**.** *The transition density of the linear differential equation* (2.83) *with* $\mathbf{u}(t) = \mathbf{0}$ *can be written in form*

$$p(\mathbf{x}(t_{k+1}) \,|\, \mathbf{x}(t_k)) = \mathrm{N}(\mathbf{x}(t_{k+1}) \,|\, \mathbf{A}_k\,\mathbf{x}(t_k), \mathbf{Q}_k), \tag{2.95}$$

where the matrices $\mathbf{A}_k$ *and* $\mathbf{Q}_k$ *are the solutions* $\mathbf{A}_k \triangleq \mathbf{A}(t_{k+1})$ *and* $\mathbf{Q}_k \triangleq \mathbf{Q}(t_{k+1})$ *to the differential equations*

$$\frac{\mathrm{d}\mathbf{A}(t)}{\mathrm{d}t} = \mathbf{F}(t)\,\mathbf{A}(t) \tag{2.96}$$

$$\frac{\mathrm{d}\mathbf{Q}(t)}{\mathrm{d}t} = \mathbf{F}(t)\,\mathbf{Q}(t) + \mathbf{Q}(t)\,\mathbf{F}^T(t) + \mathbf{L}(t)\,\mathbf{Q}_c(t)\,\mathbf{L}^T(t), \tag{2.97}$$

with the initial conditions $\mathbf{A}(t_k) = \mathbf{I}$ *and* $\mathbf{Q}(t_k) = \mathbf{0}$*. The mean and covariance of the Gaussian process solution to the equation* (2.83) *at discrete time instances* $t_1, t_2, \ldots$ *are exactly given by the recursion equations*

$$\mathbf{m}_{k+1} = \mathbf{A}_k\,\mathbf{m}_k \tag{2.98}$$

$$\mathbf{P}_{k+1} = \mathbf{A}_k\,\mathbf{P}_k\,\mathbf{A}_k^T + \mathbf{Q}_k, \tag{2.99}$$

where $\mathbf{m}_k \triangleq \mathbf{m}(t_k)$ *and* $\mathbf{P}_k \triangleq \mathbf{P}(t_k)$*.*

Theorem 2.9 (Discretization of LTI SDE)**.** *In LTI case* (2.87) *the discretization equations can be explicitly solved:*

$$\mathbf{A}_k = \exp(\mathbf{F}\,\Delta t_k) \tag{2.100}$$

$$\mathbf{Q}_k = \int_0^{\Delta t_k} \exp(\mathbf{F}\,(\Delta t_k - \tau))\,\mathbf{L}\,\mathbf{Q}_c\,\mathbf{L}^T\,\exp(\mathbf{F}\,(\Delta t_k - \tau))^T \mathrm{d}\tau, \tag{2.101}$$

where $\Delta t_k = t_{k+1} - t_k$. The matrix $\mathbf{Q}_k$ can be efficiently computed by the matrix fraction decomposition if the integral (2.101) *cannot be computed in closed form.*

The idea of discretization above is particularly useful in the case of the Kalman filter (Kalman, 1960b), because the canonical form of the Kalman filter has this kind of discrete dynamic model. The conclusion is that it does not matter that the Kalman filter was originally designed for discrete models, it still is exact for linear continuous-time dynamical models with discrete measurements.

Example 2.3 (Discretized Wiener velocity model)**.** *In Wiener velocity model (see, e.g., Bar-Shalom et al., 2001) the velocity (the first derivative of the process) is modeled as a Wiener process, that is, as a Brownian motion. In white noise interpretation this means that the acceleration (i.e., the second derivative) is a white noise process with spectral density q:*

$$\frac{\mathrm{d}^2 x(t)}{\mathrm{d}t^2} = w(t). \tag{2.102}$$

In more rigorous state space form this model can be written as

$$\begin{pmatrix} \mathrm{d}x_1 \\ \mathrm{d}x_2 \end{pmatrix} = \underbrace{\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}}_{\mathbf{F}} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mathrm{d}t + \underbrace{\begin{pmatrix} 0 \\ 1 \end{pmatrix}}_{\mathbf{L}} \mathrm{d}\beta(t), \tag{2.103}$$

where $\beta(t)$ is a Brownian motion with diffusion coefficient q, where $x_1(t) \triangleq x(t)$ is the actual process and $x_2(t)$ is its derivative. Now the matrices of the equivalent discrete-time model are given as follows:

$$\begin{aligned} \mathbf{A}(\Delta t) &= \begin{pmatrix} 1 & \Delta t \\ 0 & 1 \end{pmatrix} \\ \mathbf{Q}(\Delta t) &= \begin{pmatrix} \frac{1}{3}\Delta t^3 & \frac{1}{2}\Delta t^2 \\ \frac{1}{2}\Delta t^2 & \Delta t \end{pmatrix} q. \end{aligned} \tag{2.104}$$

2.2.7 Monte Carlo Simulation of SDEs

Assume that we are interested in forming a Monte Carlo approximation to the probability density of the state $\mathbf{x}(t)$, which is generated by the stochastic differential equation

$$\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\,\mathrm{d}t + \mathbf{L}(\mathbf{x}, t)\,\mathrm{d}\boldsymbol{\beta}. \tag{2.105}$$

where $\boldsymbol{\beta}(t)$ is a Brownian motion with diffusion matrix $\mathbf{Q}_c(t)$.

Perhaps the simplest algorithm for simulation of stochastic differential equations is the Euler-Maruyama method (see, e.g., Kloeden and Platen, 1999):

Algorithm 2.1 (Euler-Maruyama method)**.** *Draw $\mathbf{x}_0 \sim p(\mathbf{x}_0)$ and divide time $[0, t]$ interval into K steps of length Δt. On each step k do the following:*

1. *Draw random variable $\Delta\boldsymbol{\beta}_k$ from the distribution (where $t_k = k\,\Delta t$)*
$$\Delta\boldsymbol{\beta}_k \sim \mathrm{N}(\mathbf{0}, \mathbf{Q}(t_k)\,\Delta t). \tag{2.106}$$
2. *Compute*
$$\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{f}(\mathbf{x}_k, t_k)\,\Delta t + \mathbf{L}(\mathbf{x}_k, t_k)\Delta\boldsymbol{\beta}_k. \tag{2.107}$$

The disadvantage of this method is that although the deterministic Euler algorithm is of order $O(\Delta t)$ of convergence this stochastic algorithm is actually only of strong order $O(\Delta t^{1/2})$ (Kloeden and Platen, 1999). It still is of the weak order $O(\Delta t)$, but the strong order of convergence is what counts, for example, in the case of numerically evaluating the importance weights in the next section.

Fortunately, it is possible to modify any higher order numerical integration method for deterministic differential equations such that its strong order is half the deterministic order (Wilkie, 2004). For example, the stochastic weak fourth order, strong second order Runge-Kutta method can be implemented by defining the function

$$\hat{f}_j(\mathbf{x}, t, \Delta\boldsymbol{\beta}) = f_j(\mathbf{x}, t) - \frac{1}{2}\sum_{ik} L_{ik}(\mathbf{x}, t)\frac{\partial L_{jk}(\mathbf{x}, t)}{\partial x_i} + \sum_k L_{jk}(\mathbf{x}, t)\Delta\boldsymbol{\beta}_k/\Delta t. \tag{2.108}$$

Now the algorithm can be implemented as follows:

Algorithm 2.2 (Stochastic Runge-Kutta method)**.** *On each step k do the following:*

1. *Draw random variable $\Delta\boldsymbol{\beta}_k$ from the distribution ($t_k = k\,\Delta t$)*
$$\Delta\boldsymbol{\beta}_k \sim \mathrm{N}(\mathbf{0}, \mathbf{Q}(t_k)\,\Delta t). \tag{2.109}$$
2. *Compute*
$$\begin{aligned}
\Delta\mathbf{x}^1 &= \hat{\mathbf{f}}(\mathbf{x}_k, t_k, \Delta\boldsymbol{\beta}_k)\,\Delta t \\
\Delta\mathbf{x}^2 &= \hat{\mathbf{f}}(\mathbf{x}_k + \Delta\mathbf{x}^1/2, t_k + \Delta t/2, \Delta\boldsymbol{\beta}_k)\,\Delta t \\
\Delta\mathbf{x}^3 &= \hat{\mathbf{f}}(\mathbf{x}_k + \Delta\mathbf{x}^2/2, t_k + \Delta t/2, \Delta\boldsymbol{\beta}_k)\,\Delta t \\
\Delta\mathbf{x}^4 &= \hat{\mathbf{f}}(\mathbf{x}_k + \Delta\mathbf{x}^3, t_k + \Delta t, \Delta\boldsymbol{\beta}_k)\,\Delta t \\
\mathbf{x}_{k+1} &= \mathbf{x}_k + \frac{1}{6}(\Delta\mathbf{x}^1 + 2\Delta\mathbf{x}^2 + 2\Delta\mathbf{x}^3 + \Delta\mathbf{x}^4).
\end{aligned} \tag{2.110}$$

The idea of this algorithm is that the Itô SDE is actually converted into the corresponding Stratonovich differential equation (hence the correction term in $\hat{\mathbf{f}}$). The fortunate property of this Stratonovich form is that the Taylor series for functions can be formed in the same way as in deterministic case. Thus the Runge-Kutta method can be derived in the same way as in deterministic case, but now the strong order is half, because $\Delta\beta^2$ is of the order $O(\Delta t)$.

Chapter 3

Optimal Filtering and Smoothing

This chapter presents the theory and relevant algorithms of discrete-time filtering and smoothing, the theory and relevant algorithms of continuous-discrete filtering and smoothing, and Gaussian approximation based algorithms for continuous-time filtering and smoothing. Most of the methodological and theoretical contributions of this thesis can be found in this chapter:

- The *matrix form of the unscented transform* is presented in Section 3.1.
- The *continuous-discrete unscented Kalman filter* and the *continuous-discrete unscented Kalman smoother* are presented in Section 3.2.
- Measure transformation based *continuous-discrete particle filters* and *particle smoothers* are presented in Section 3.2.
- The continuous-time unscented Kalman filter, the *unscented Kalman-Bucy filter* is presented in Section 3.3.

3.1 Discrete-Time Filtering and Smoothing

This section first presents the classical formulation of the discrete-time optimal filtering as recursive Bayesian inference. Then the classical (extended) Kalman filters and smoothers are presented in terms of the general theory. In addition to the classical algorithms the unscented Kalman filter and unscented Kalman smoother are compactly formulated in terms of the matrix form unscented transform. Sequential importance resampling, as well as Rao-Blackwellized particle filtering and smoothing are also covered.

3.1.1 Discrete-Time Filtering and Smoothing Equations

Before going into the practical non-linear filtering algorithms, in the next sections the theory of probabilistic (Bayesian) filtering is presented. The Kalman filtering and smoothing equations, which are the closed form solutions to the linear Gaussian discrete-time optimal filtering problem, are also derived.

Discrete-Time State Space Models

Definition 3.1 (Discrete-time state space model)**.** Discrete-time state space model *is a recursively defined probabilistic model of the form*

$$\begin{aligned} \mathbf{x}_k &\sim p(\mathbf{x}_k \,|\, \mathbf{x}_{k-1}) \\ \mathbf{y}_k &\sim p(\mathbf{y}_k \,|\, \mathbf{x}_k), \end{aligned} \tag{3.1}$$

where

- $\mathbf{x}_k \in \mathbb{R}^n$ *is the* state *of the system on the time step* k*.*
- $\mathbf{y}_k \in \mathbb{R}^m$ *is the measurement on the time step* k*.*
- $p(\mathbf{x}_k \,|\, \mathbf{x}_{k-1})$ *is the* dynamic model*, which models the stochastic dynamics of the system. The dynamic model can be a probability density, a counting measure or combination of them depending on if the state* $\mathbf{x}_k$ *is continuous, discrete or hybrid.*
- $p(\mathbf{y}_k \,|\, \mathbf{x}_k)$ *is the* measurement model*, which models the distribution of the measurements given the state.*

The model has the following properties (i.e., assumptions):

Property 3.1 (Markov property of states)**.**

States $\{\mathbf{x}_k \,:\, k = 1, 2, \ldots\}$ *form a Markov sequence (or Markov chain if the state is discrete). This Markov property means that* $\mathbf{x}_k$ *(and actually the whole future* $\mathbf{x}_{k+1}, \mathbf{x}_{k+2}, \ldots$*) given* $\mathbf{x}_{k-1}$ *is independent from anything that has happened in the past:*

$$p(\mathbf{x}_k \,|\, \mathbf{x}_{1:k-1}, \mathbf{y}_{1:k-1}) = p(\mathbf{x}_k \,|\, \mathbf{x}_{k-1}). \tag{3.2}$$

Also the past is independent of the future given the present:

$$p(\mathbf{x}_{k-1} \,|\, \mathbf{x}_{k:T}, \mathbf{y}_{k:T}) = p(\mathbf{x}_{k-1} \,|\, \mathbf{x}_k). \tag{3.3}$$

Property 3.2 (Conditional independence of measurements)**.**

The measurement $\mathbf{y}_k$ *given the* $\mathbf{x}_k$ *is conditionally independent from the measurement and state histories:*

$$p(\mathbf{y}_k \,|\, \mathbf{x}_{1:k}, \mathbf{y}_{1:k-1}) = p(\mathbf{y}_k \,|\, \mathbf{x}_k). \tag{3.4}$$

Example 3.1 (Gaussian random walk)**.** *Gaussian random walk model can be written as*

$$\begin{aligned} x_k &= x_{k-1} + w_{k-1}, \quad & w_{k-1} &\sim \mathrm{N}(0, q) \\ y_k &= x_k + e_k, & e_k &\sim \mathrm{N}(0, r), \end{aligned} \tag{3.5}$$

where x_k is the hidden state and y_k is the measurement. In terms of probability densities the model can be written as

$$\begin{aligned} p(x_k \,|\, x_{k-1}) &= \mathrm{N}(x_k \,|\, x_{k-1}, q) \\ &= \frac{1}{\sqrt{2\pi q}} \exp\left(-\frac{1}{2q} (x_k - x_{k-1})^2 \right) \\ p(y_k \,|\, x_k) &= \mathrm{N}(y_k \,|\, x_k, r) \\ &= \frac{1}{\sqrt{2\pi r}} \exp\left(-\frac{1}{2r} (y_k - x_k)^2 \right) \end{aligned} \tag{3.6}$$

which is a discrete-time state space model.

The filtering model actually states that the joint prior distribution of the states $(\mathbf{x}_0, \ldots, \mathbf{x}_T)$ and the joint likelihood of the measurements $(\mathbf{y}_0, \ldots, \mathbf{y}_T)$ are, respectively

$$p(\mathbf{x}_0, \ldots, \mathbf{x}_T) = p(\mathbf{x}_0) \prod_{k=1}^{T} p(\mathbf{x}_k \,|\, \mathbf{x}_{k-1}) \tag{3.7}$$

$$p(\mathbf{y}_1, \ldots, \mathbf{y}_T \,|\, \mathbf{x}_0, \ldots, \mathbf{x}_T) = \prod_{k=1}^{T} p(\mathbf{y}_k \,|\, \mathbf{x}_k). \tag{3.8}$$

In principle, for given T we could simply compute the posterior distribution of the states by the Bayes rule:

$$\begin{aligned} p(\mathbf{x}_0, \ldots, \mathbf{x}_T \,|\, \mathbf{y}_1, \ldots, \mathbf{y}_T) &= \frac{p(\mathbf{y}_1, \ldots, \mathbf{y}_T \,|\, \mathbf{x}_0, \ldots, \mathbf{x}_T) \, p(\mathbf{x}_0, \ldots, \mathbf{x}_T)}{p(\mathbf{y}_1, \ldots, \mathbf{y}_T)} \\ &\propto p(\mathbf{y}_1, \ldots, \mathbf{y}_T \,|\, \mathbf{x}_0, \ldots, \mathbf{x}_T) \, p(\mathbf{x}_0, \ldots, \mathbf{x}_T). \end{aligned} \tag{3.9}$$

However, this kind of explicit computation is not feasible in real time applications, because the amount of computations per time step increases when new observations arrive. Thus, this way we could only work with small data sets, because if the amount of data is not bounded (as in real time sensoring applications), at some point of time the computations will become intractable. To cope with real time data we need to have algorithm where we do constant amount of computations per time step.

Due to the problem formulation, instead of the full joint posterior distribution of the states more useful distributions are the *filtering distributions* and *smoothing distributions*, which shall be discussed next.

Discrete-Time Optimal Filtering Equations

The purpose of *optimal filtering* is to compute the *marginal posterior distribution* of the state $\mathbf{x}_k$ on the time step k given the history of the measurements up to the time step k

$$p(\mathbf{x}_k \,|\, \mathbf{y}_{1:k}). \tag{3.10}$$

The fundamental equations of the Bayesian filtering theory are given by the following theorem:

Theorem 3.1 (Discrete-time Bayesian filtering equations)**.** *The recursive equations for computing the* predicted distribution $p(\mathbf{x}_k \,|\, \mathbf{y}_{1:k-1})$ *and the* filtering distribution $p(\mathbf{x}_k \,|\, \mathbf{y}_{1:k})$ *on the time step k are given by the following* Bayesian filtering equations*:*

- Initialization. *The recursion starts from the prior distribution $p(\mathbf{x}_0)$.*
- Prediction. *The predictive distribution of the state $\mathbf{x}_k$ on time step k given the dynamic model can be computed by the Chapman-Kolmogorov equation*

$$p(\mathbf{x}_k \,|\, \mathbf{y}_{1:k-1}) = \int p(\mathbf{x}_k \,|\, \mathbf{x}_{k-1})\, p(\mathbf{x}_{k-1} \,|\, \mathbf{y}_{1:k-1})\, \mathrm{d}\mathbf{x}_{k-1}. \tag{3.11}$$

- Update. *Given the measurement $\mathbf{y}_k$ on time step k the posterior distribution of the state $\mathbf{x}_k$ can be computed by the Bayes' rule*

$$p(\mathbf{x}_k \,|\, \mathbf{y}_{1:k}) = \frac{1}{Z_k} p(\mathbf{y}_k \,|\, \mathbf{x}_k)\, p(\mathbf{x}_k \,|\, \mathbf{y}_{1:k-1}), \tag{3.12}$$

where the normalization constant Z_k is given as

$$Z_k = \int p(\mathbf{y}_k \,|\, \mathbf{x}_k)\, p(\mathbf{x}_k \,|\, \mathbf{y}_{1:k-1})\, \mathrm{d}\mathbf{x}_k. \tag{3.13}$$

If some of the components of the state are discrete, the corresponding integrals are replaced with summations.

Proof. The joint distribution of $\mathbf{x}_k$ and $\mathbf{x}_{k-1}$ given $\mathbf{y}_{1:k-1}$ can be computed as

$$\begin{aligned} p(\mathbf{x}_k, \mathbf{x}_{k-1} \,|\, \mathbf{y}_{1:k-1}) &= p(\mathbf{x}_k \,|\, \mathbf{x}_{k-1}, \mathbf{y}_{1:k-1})\, p(\mathbf{x}_{k-1} \,|\, \mathbf{y}_{1:k-1}) \\ &= p(\mathbf{x}_k \,|\, \mathbf{x}_{k-1})\, p(\mathbf{x}_{k-1} \,|\, \mathbf{y}_{1:k-1}), \end{aligned} \tag{3.14}$$

where the disappearance of the measurement history $\mathbf{y}_{1:k-1}$ is due to the Markov property of the sequence $\{\mathbf{x}_k, k = 1, 2, \ldots\}$. The marginal distribution of $\mathbf{x}_k$ given $\mathbf{y}_{1:k-1}$ can be obtained by integrating the distribution (3.14) over $\mathbf{x}_{k-1}$, which gives the *Chapman-Kolmogorov equation*

$$p(\mathbf{x}_k \,|\, \mathbf{y}_{1:k-1}) = \int p(\mathbf{x}_k \,|\, \mathbf{x}_{k-1})\, p(\mathbf{x}_{k-1} \,|\, \mathbf{y}_{1:k-1})\, \mathrm{d}\mathbf{x}_{k-1}. \tag{3.15}$$

If $\mathbf{x}_{k-1}$ is discrete, then the above integral is replaced with sum over $\mathbf{x}_{k-1}$. The distribution of $\mathbf{x}_k$ given $\mathbf{y}_k$ and $\mathbf{y}_{1:k-1}$, that is, given $\mathbf{y}_{1:k}$ can be computed by the *Bayes' rule*

$$\begin{aligned} p(\mathbf{x}_k \,|\, \mathbf{y}_{1:k}) &= \frac{1}{Z_k} p(\mathbf{y}_k \,|\, \mathbf{x}_k, \mathbf{y}_{1:k-1})\, p(\mathbf{x}_k \,|\, \mathbf{y}_{1:k-1}) \\ &= \frac{1}{Z_k} p(\mathbf{y}_k \,|\, \mathbf{x}_k)\, p(\mathbf{x}_k \,|\, \mathbf{y}_{1:k-1}) \end{aligned} \tag{3.16}$$

where the normalization constant is given by Equation (3.13). The disappearance of the measurement history $\mathbf{y}_{1:k-1}$ in the Equation (3.16) is due to the conditional independence of $\mathbf{y}_k$ from the measurement history, given $\mathbf{x}_k$. □

Discrete-Time Optimal Smoothing Equations

The purpose of *optimal smoothing*[1] is to compute the marginal posterior distribution of the state $\mathbf{x}_k$ at the time step k after receiving the measurements up to a time step T, where $T > k$:

$$p(\mathbf{x}_k \,|\, \mathbf{y}_{1:T}). \tag{3.17}$$

The difference between filters and smoothers is that *the optimal filter* computes its estimates using only the measurements obtained before and on the time step k, but *the optimal smoother* uses also the future measurements for computing its estimates. After obtaining the filtering posterior state distributions, the following theorem gives the equations for computing the marginal posterior distributions for each time step conditionally to all measurements up to the time step T:

Theorem 3.2 (Discrete-time Bayesian fixed interval smoother)**.** *The backward recursive equations for computing the* smoothed distributions $p(\mathbf{x}_k \,|\, \mathbf{y}_{1:T})$ *for any* $k < T$ *are given by the following* Bayesian (fixed interval) smoothing equations

$$\begin{aligned} p(\mathbf{x}_{k+1} \,|\, \mathbf{y}_{1:k}) &= \int p(\mathbf{x}_{k+1} \,|\, \mathbf{x}_k)\, p(\mathbf{x}_k \,|\, \mathbf{y}_{1:k})\, \mathrm{d}\mathbf{x}_k \\ p(\mathbf{x}_k \,|\, \mathbf{y}_{1:T}) &= p(\mathbf{x}_k \,|\, \mathbf{y}_{1:k}) \int \left[\frac{p(\mathbf{x}_{k+1} \,|\, \mathbf{x}_k)\, p(\mathbf{x}_{k+1} \,|\, \mathbf{y}_{1:T})}{p(\mathbf{x}_{k+1} \,|\, \mathbf{y}_{1:k})} \right] \mathrm{d}\mathbf{x}_{k+1}, \end{aligned} \tag{3.18}$$

where $p(\mathbf{x}_k \,|\, \mathbf{y}_{1:k})$ *is the filtering distribution of the time step* k*. Note that the term* $p(\mathbf{x}_{k+1} \,|\, \mathbf{y}_{1:k})$ *is simply the predicted distribution of time step* $k+1$*. The integrations are replaced by summations if some of the state components are discrete.*

Proof. Due to the Markov properties the state $\mathbf{x}_k$ is independent of $\mathbf{y}_{k+1:T}$ given $\mathbf{x}_{k+1}$, which gives $p(\mathbf{x}_k \,|\, \mathbf{x}_{k+1}, \mathbf{y}_{1:T}) = p(\mathbf{x}_k \,|\, \mathbf{x}_{k+1}, \mathbf{y}_{1:k})$. By using the *Bayes'*

[1] In this thesis only fixed-interval smoothing is considered.

rule the distribution of $\mathbf{x}_k$ given $\mathbf{x}_{k+1}$ and $\mathbf{y}_{1:T}$ can be expressed as

$$\begin{aligned} p(\mathbf{x}_k \,|\, \mathbf{x}_{k+1}, \mathbf{y}_{1:T}) &= p(\mathbf{x}_k \,|\, \mathbf{x}_{k+1}, \mathbf{y}_{1:k}) \\ &= \frac{p(\mathbf{x}_k, \mathbf{x}_{k+1} \,|\, \mathbf{y}_{1:k})}{p(\mathbf{x}_{k+1} \,|\, \mathbf{y}_{1:k})} \\ &= \frac{p(\mathbf{x}_{k+1} \,|\, \mathbf{x}_k, \mathbf{y}_{1:k})\, p(\mathbf{x}_k \,|\, \mathbf{y}_{1:k})}{p(\mathbf{x}_{k+1} \,|\, \mathbf{y}_{1:k})} \\ &= \frac{p(\mathbf{x}_{k+1} \,|\, \mathbf{x}_k)\, p(\mathbf{x}_k \,|\, \mathbf{y}_{1:k})}{p(\mathbf{x}_{k+1} \,|\, \mathbf{y}_{1:k})}. \end{aligned} \tag{3.19}$$

The joint distribution of $\mathbf{x}_k$ and $\mathbf{x}_{k+1}$ given $\mathbf{y}_{1:T}$ can be now computed as

$$\begin{aligned} p(\mathbf{x}_k, \mathbf{x}_{k+1} \,|\, \mathbf{y}_{1:T}) &= p(\mathbf{x}_k \,|\, \mathbf{x}_{k+1}, \mathbf{y}_{1:T})\, p(\mathbf{x}_{k+1} \,|\, \mathbf{y}_{1:T}) \\ &= p(\mathbf{x}_k \,|\, \mathbf{x}_{k+1}, \mathbf{y}_{1:k})\, p(\mathbf{x}_{k+1} \,|\, \mathbf{y}_{1:T}) \\ &= \frac{p(\mathbf{x}_{k+1} \,|\, \mathbf{x}_k)\, p(\mathbf{x}_k \,|\, \mathbf{y}_{1:k})\, p(\mathbf{x}_{k+1} \,|\mathbf{y}_{1:T})}{p(\mathbf{x}_{k+1} \,|\, \mathbf{y}_{1:k})}, \end{aligned} \tag{3.20}$$

where $p(\mathbf{x}_{k+1} \,|\, \mathbf{y}_{1:T})$ is the smoothed distribution of the time step $k+1$. The marginal distribution of $\mathbf{x}_k$ given $\mathbf{y}_{1:T}$ is given by integral (or summation) over $\mathbf{x}_{k+1}$ in Equation (3.20), which gives the desired result. □

Discrete-Time Kalman Filter

The discrete-time Kalman filter (Kalman, 1960b) is the closed form solution to the optimal filtering equations of the discrete-time filtering model, where the dynamic and measurements models are linear Gaussian:

$$\begin{aligned} \mathbf{x}_k &= \mathbf{A}_{k-1}\, \mathbf{x}_{k-1} + \mathbf{q}_{k-1} \\ \mathbf{y}_k &= \mathbf{H}_k\, \mathbf{x}_k + \mathbf{r}_k, \end{aligned} \tag{3.21}$$

where $\mathbf{x}_k \in \mathbb{R}^n$ is the state, $\mathbf{y}_k \in \mathbb{R}^m$ is the measurement, $\mathbf{q}_{k-1} \sim \mathrm{N}(\mathbf{0}, \mathbf{Q}_{k-1})$ is the process noise, $\mathbf{r}_k \sim \mathrm{N}(\mathbf{0}, \mathbf{R}_k)$ is the measurement noise and the prior distribution is Gaussian $\mathbf{x}_0 \sim \mathrm{N}(\mathbf{m}_0, \mathbf{P}_0)$. The matrix $\mathbf{A}_{k-1}$ is the transition matrix of the dynamic model and $\mathbf{H}_k$ is the measurement model matrix. In probabilistic terms the model is

$$\begin{aligned} p(\mathbf{x}_k \,|\, \mathbf{x}_{k-1}) &= \mathrm{N}(\mathbf{x}_k \,|\, \mathbf{A}_{k-1}\, \mathbf{x}_{k-1}, \mathbf{Q}_{k-1}) \\ p(\mathbf{y}_k \,|\, \mathbf{x}_k) &= \mathrm{N}(\mathbf{y}_k \,|\, \mathbf{H}_k\, \mathbf{x}_k, \mathbf{R}_k). \end{aligned} \tag{3.22}$$

Theorem 3.3 (Discrete-time Kalman filter)**.** *The optimal filtering equations for the linear filtering model* (3.21) *can be evaluated in closed form and the resulting distributions are Gaussian:*

$$\begin{aligned} p(\mathbf{x}_k \,|\, \mathbf{y}_{1:k-1}) &= \mathrm{N}(\mathbf{x}_k \,|\, \mathbf{m}_k^-, \mathbf{P}_k^-) \\ p(\mathbf{x}_k \,|\, \mathbf{y}_{1:k}) &= \mathrm{N}(\mathbf{x}_k \,|\, \mathbf{m}_k, \mathbf{P}_k) \\ p(\mathbf{y}_k \,|\, \mathbf{y}_{1:k-1}) &= \mathrm{N}(\mathbf{y}_k \,|\, \mathbf{H}_k \mathbf{m}_k^-, \mathbf{S}_k). \end{aligned} \tag{3.23}$$

The parameters of the distributions above can be computed with the following discrete Kalman filter prediction *and* update steps*:*

- The prediction step *is*

$$\begin{aligned} \mathbf{m}_k^- &= \mathbf{A}_{k-1}\,\mathbf{m}_{k-1} \\ \mathbf{P}_k^- &= \mathbf{A}_{k-1}\,\mathbf{P}_{k-1}\,\mathbf{A}_{k-1}^T + \mathbf{Q}_{k-1}. \end{aligned} \tag{3.24}$$

- The update step *is*

$$\begin{aligned} \mathbf{v}_k &= \mathbf{y}_k - \mathbf{H}_k\,\mathbf{m}_k^- \\ \mathbf{S}_k &= \mathbf{H}_k\,\mathbf{P}_k^-\,\mathbf{H}_k^T + \mathbf{R}_k \\ \mathbf{K}_k &= \mathbf{P}_k^-\,\mathbf{H}_k^T\,\mathbf{S}_k^{-1} \\ \mathbf{m}_k &= \mathbf{m}_k^- + \mathbf{K}_k\,\mathbf{v}_k \\ \mathbf{P}_k &= \mathbf{P}_k^- - \mathbf{K}_k\,\mathbf{S}_k\,\mathbf{K}_k^T. \end{aligned} \tag{3.25}$$

Proof. By Lemma A.1 on page 207, the joint distribution of $\mathbf{x}_k$ and $\mathbf{x}_{k-1}$ given $\mathbf{y}_{1:k-1}$ is

$$\begin{aligned} p(\mathbf{x}_{k-1}, \mathbf{x}_k \,|\, \mathbf{y}_{1:k-1}) &= p(\mathbf{x}_k \,|\, \mathbf{x}_{k-1})\, p(\mathbf{x}_{k-1} \,|\, \mathbf{y}_{1:k-1}) \\ &= \mathrm{N}(\mathbf{x}_k \,|\, \mathbf{A}_{k-1}\,\mathbf{x}_{k-1}, \mathbf{Q}_{k-1})\ \mathrm{N}(\mathbf{x}_{k-1} \,|\, \mathbf{m}_{k-1}, \mathbf{P}_{k-1}) \\ &= \mathrm{N}\left(\begin{bmatrix} \mathbf{x}_{k-1} \\ \mathbf{x}_k \end{bmatrix} \,\middle|\, \mathbf{m}_1, \mathbf{P}_1 \right), \end{aligned} \tag{3.26}$$

where

$$\mathbf{m}_1 = \begin{pmatrix} \mathbf{m}_{k-1} \\ \mathbf{A}_{k-1}\,\mathbf{m}_{k-1} \end{pmatrix}, \quad \mathbf{P}_1 = \begin{pmatrix} \mathbf{P}_{k-1} & \mathbf{P}_{k-1}\,\mathbf{A}_{k-1}^T \\ \mathbf{A}_{k-1}\,\mathbf{P}_{k-1} & \mathbf{A}_{k-1}\,\mathbf{P}_{k-1}\,\mathbf{A}_{k-1}^T + \mathbf{Q}_{k-1} \end{pmatrix}. \tag{3.27}$$

and the marginal distribution of $\mathbf{x}_k$ is by Lemma A.2

$$p(\mathbf{x}_k \,|\, \mathbf{y}_{1:k-1}) = \mathrm{N}(\mathbf{x}_k \,|\, \mathbf{m}_k^-, \mathbf{P}_k^-), \tag{3.28}$$

where

$$\mathbf{m}_k^- = \mathbf{A}_{k-1}\,\mathbf{m}_{k-1}, \qquad \mathbf{P}_k^- = \mathbf{A}_{k-1}\,\mathbf{P}_{k-1}\,\mathbf{A}_{k-1}^T + \mathbf{Q}_{k-1}. \tag{3.29}$$

By Lemma A.1, the joint distribution of $\mathbf{y}_k$ and $\mathbf{x}_k$ is

$$\begin{aligned} p(\mathbf{x}_k, \mathbf{y}_k \,|\, \mathbf{y}_{1:k-1}) &= p(\mathbf{y}_k \,|\, \mathbf{x}_k)\, p(\mathbf{x}_k \,|\, \mathbf{y}_{1:k-1}) \\ &= \mathrm{N}(\mathbf{y}_k \,|\, \mathbf{H}_k\,\mathbf{x}_k, \mathbf{R}_k)\ \mathrm{N}(\mathbf{x}_k \,|\, \mathbf{m}_k^-, \mathbf{P}_k^-) \\ &= \mathrm{N}\left(\begin{bmatrix} \mathbf{x}_k \\ \mathbf{y}_k \end{bmatrix} \,\middle|\, \mathbf{m}_2, \mathbf{P}_2 \right), \end{aligned} \tag{3.30}$$

where

$$\mathbf{m}_2 = \begin{pmatrix} \mathbf{m}_k^- \\ \mathbf{H}_k \, \mathbf{m}_k^- \end{pmatrix}, \qquad \mathbf{P}_2 = \begin{pmatrix} \mathbf{P}_k^- & \mathbf{P}_k^- \, \mathbf{H}_k^T \\ \mathbf{H}_k \, \mathbf{P}_k^- & \mathbf{H}_k \, \mathbf{P}_k^- \, \mathbf{H}_k^T + \mathbf{R}_k \end{pmatrix}. \tag{3.31}$$

By Lemma A.2 the conditional distribution of $\mathbf{x}_k$ is

$$\begin{aligned} p(\mathbf{x}_k \,|\, \mathbf{y}_k, \mathbf{y}_{1:k-1}) &= p(\mathbf{x}_k \,|\, \mathbf{y}_{1:k}) \\ &= \mathrm{N}(\mathbf{x}_k \,|\, \mathbf{m}_k, \mathbf{P}_k), \end{aligned} \tag{3.32}$$

where

$$\begin{aligned} \mathbf{S}_k &= \mathbf{H}_k \, \mathbf{P}_k^- \, \mathbf{H}_k^T + \mathbf{R}_k \\ \mathbf{K}_k &= \mathbf{P}_k^- \, \mathbf{H}_k^T \, \mathbf{S}_k^{-1} \\ \mathbf{m}_k &= \mathbf{m}_k^- + \mathbf{K}_k \, [\mathbf{y}_k - \mathbf{H}_k \, \mathbf{m}_k^-] \\ \mathbf{P}_k &= \mathbf{P}_k^- - \mathbf{K}_k \, \mathbf{S}_k \, \mathbf{K}_k^T. \end{aligned} \tag{3.33}$$

□

The functional form of the Kalman filter equations given here is not the only possible one. In the numerical stability point of view it would be better to work with matrix square roots of covariances instead of plain covariance matrices. The theory and details of implementation of this kind of methods is well covered, for example, in the book of Grewal and Andrews (2001).

Example 3.2 (Kalman filter for Gaussian random walk)**.** *Assume that we are observing measurements y_k of the Gaussian random walk model given in Example 3.1 and we want to estimate the state x_k on each time step. The information obtained up to time step $k-1$ is summarized by the Gaussian filtering density*

$$p(x_{k-1} \,|\, y_{1:k-1}) = \mathrm{N}(x_{k-1} \,|\, m_{k-1}, P_{k-1}). \tag{3.34}$$

The Kalman filter prediction and update equations are now given as

$$\begin{aligned} m_k^- &= m_{k-1} \\ P_k^- &= P_{k-1} + q \\ m_k &= m_k^- + \frac{P_k^-}{P_k^- + r}(y_k - m_k^-) \\ P_k &= P_k^- - \frac{(P_k^-)^2}{P_k^- + r}. \end{aligned} \tag{3.35}$$

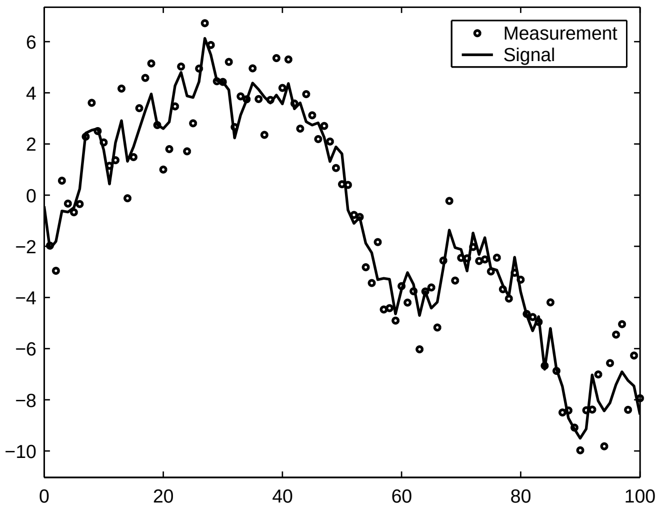

Figure 3.1: Simulated signal and measurements of the Kalman filtering example (Example 3.2).

Discrete-Time Kalman Smoother

The *discrete-time Kalman smoother*[2] (see, e.g., Rauch et al., 1965; Gelb, 1974; Bar-Shalom et al., 2001) can be used for computing the closed form smoothing solution

$$p(\mathbf{x}_k \,|\, \mathbf{y}_{1:T}) = \mathrm{N}(\mathbf{x}_k \,|\, \mathbf{m}_k^s, \mathbf{P}_k^s), \tag{3.36}$$

to the linear filtering model (3.21). The difference to the solution computed by the *Kalman filter* is that the smoothed solution is conditional on the whole measurement data $\mathbf{y}_{1:T}$, while the filtering solution is conditional only on the measurements obtained before and on the time step k, that is, on the measurements $\mathbf{y}_{1:k}$.

Theorem 3.4 (Discrete-time Kalman smoother)**.** *The backward recursion equations for the discrete-time fixed interval Kalman smoother (Rauch-Tung-Striebel smoother) are given as*

$$\begin{aligned}
\mathbf{m}_{k+1}^- &= \mathbf{A}_k \, \mathbf{m}_k \\
\mathbf{P}_{k+1}^- &= \mathbf{A}_k \, \mathbf{P}_k \, \mathbf{A}_k^T + \mathbf{Q}_k \\
\mathbf{C}_k &= \mathbf{P}_k \, \mathbf{A}_k^T \, [\mathbf{P}_{k+1}^-]^{-1} \\
\mathbf{m}_k^s &= \mathbf{m}_k + \mathbf{C}_k \, [\mathbf{m}_{k+1}^s - \mathbf{m}_{k+1}^-] \\
\mathbf{P}_k^s &= \mathbf{P}_k + \mathbf{C}_k \, [\mathbf{P}_{k+1}^s - \mathbf{P}_{k+1}^-] \, \mathbf{C}_k^T,
\end{aligned} \tag{3.37}$$

[2] Also called discrete-time Rauch-Tung-Striebel (RTS) smoother.

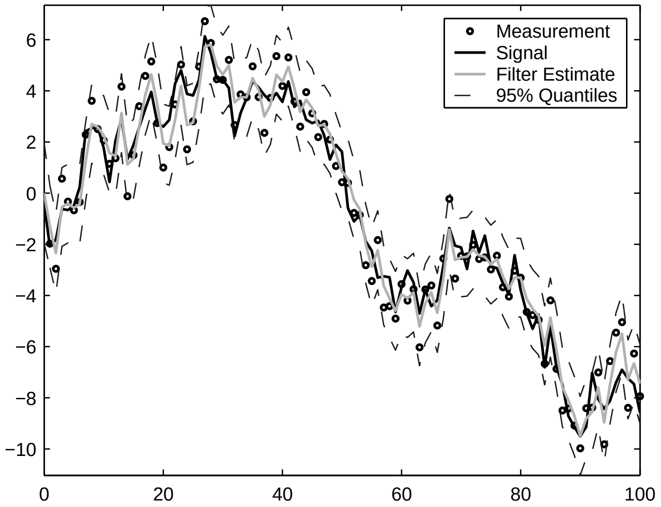

Figure 3.2: Signal, measurements and filtering estimate of the Kalman filtering example (Example 3.2).

where $\mathbf{m}_k$ *and* $\mathbf{P}_k$ *are the mean and covariance computed by the Kalman filter. The recursion is started from the last time step* T*, with* $\mathbf{m}_T^s = \mathbf{m}_T$ *and* $\mathbf{P}_T^s = \mathbf{P}_T$*. Note that the first two of the equations are simply the Kalman filter prediction equations.*

Proof. Similarly to the Kalman filter case, by Lemma A.1, the joint distribution of $\mathbf{x}_k$ and $\mathbf{x}_{k+1}$ given $\mathbf{y}_{1:k}$ is

$$\begin{aligned} p(\mathbf{x}_k, \mathbf{x}_{k+1} \,|\, \mathbf{y}_{1:k}) &= p(\mathbf{x}_{k+1} \,|\, \mathbf{x}_k)\, p(\mathbf{x}_k \,|\, \mathbf{y}_{1:k}) \\ &= \mathrm{N}(\mathbf{x}_{k+1} \,|\, \mathbf{A}_k\, \mathbf{x}_k, \mathbf{Q}_k)\ \mathrm{N}(\mathbf{x}_k \,|\, \mathbf{m}_k, \mathbf{P}_k) \\ &= \mathrm{N}\left(\begin{bmatrix} \mathbf{x}_k \\ \mathbf{x}_{k+1} \end{bmatrix} \,\middle|\, \mathbf{m}_1, \mathbf{P}_1 \right), \end{aligned} \tag{3.38}$$

where

$$\mathbf{m}_1 = \begin{pmatrix} \mathbf{m}_k \\ \mathbf{A}_k\, \mathbf{m}_k \end{pmatrix}, \qquad \mathbf{P}_1 = \begin{pmatrix} \mathbf{P}_k & \mathbf{P}_k\, \mathbf{A}_k^T \\ \mathbf{A}_k\, \mathbf{P}_k & \mathbf{A}_k\, \mathbf{P}_k\, \mathbf{A}_k^T + \mathbf{Q}_k \end{pmatrix}. \tag{3.39}$$

Due to the Markov property of the states we have

$$p(\mathbf{x}_k \,|\, \mathbf{x}_{k+1}, \mathbf{y}_{1:T}) = p(\mathbf{x}_k \,|\, \mathbf{x}_{k+1}, \mathbf{y}_{1:k}), \tag{3.40}$$

and thus by Lemma A.2 we get the conditional distribution

$$\begin{aligned} p(\mathbf{x}_k \,|\, \mathbf{x}_{k+1}, \mathbf{y}_{1:T}) &= p(\mathbf{x}_k \,|\, \mathbf{x}_{k+1}, \mathbf{y}_{1:k}) \\ &= \mathrm{N}(\mathbf{x}_k \,|\, \mathbf{m}_2, \mathbf{P}_2), \end{aligned} \tag{3.41}$$

where

$$\begin{aligned}
\mathbf{P}_{k+1}^- &= \mathbf{A}_k\,\mathbf{P}_k\,\mathbf{A}_k^T + \mathbf{Q}_k \\
\mathbf{C}_k &= \mathbf{P}_k\,\mathbf{A}_k^T\,[\mathbf{P}_{k+1}^-]^{-1} \\
\mathbf{m}_2 &= \mathbf{m}_k + \mathbf{C}_k\,(\mathbf{x}_{k+1} - \mathbf{A}_k\,\mathbf{m}_k) \\
&= \mathbf{C}_k\,\mathbf{x}_{k+1} + (\mathbf{I} - \mathbf{C}_k\,\mathbf{A}_k)\,\mathbf{m}_k \\
\mathbf{P}_2 &= \mathbf{P}_k - \mathbf{C}_k\,\mathbf{P}_{k+1}^-\,\mathbf{C}_k^T.
\end{aligned} \tag{3.42}$$

The joint distribution of $\mathbf{x}_k$ and $\mathbf{x}_{k+1}$ given all the data is

$$\begin{aligned}
p(\mathbf{x}_{k+1}, \mathbf{x}_k \mid \mathbf{y}_{1:T}) &= p(\mathbf{x}_k \mid \mathbf{x}_{k+1}, \mathbf{y}_{1:T})\,p(\mathbf{x}_{k+1} \mid \mathbf{y}_{1:T}) \\
&= \mathrm{N}(\mathbf{x}_k \mid \mathbf{m}_2, \mathbf{P}_2)\ \mathrm{N}(\mathbf{x}_{k+1} \mid \mathbf{m}_{k+1}^s, \mathbf{P}_{k+1}^s) \\
&= \mathrm{N}\left(\begin{bmatrix}\mathbf{x}_{k+1} \\ \mathbf{x}_k\end{bmatrix} \,\middle|\, \mathbf{m}_3, \mathbf{P}_3\right)
\end{aligned} \tag{3.43}$$

where

$$\begin{aligned}
\mathbf{m}_3 &= \begin{pmatrix} \mathbf{m}_{k+1}^s \\ \mathbf{C}_k\,\mathbf{x}_{k+1} + (\mathbf{I} - \mathbf{C}_k\,\mathbf{A}_k)\,\mathbf{m}_k \end{pmatrix} \\
\mathbf{P}_3 &= \begin{pmatrix} \mathbf{P}_{k+1}^s & \mathbf{P}_{k+1}^s\,\mathbf{C}_k^T \\ \mathbf{C}_k\,\mathbf{P}_{k+1}^s & \mathbf{C}_k\,\mathbf{P}_{k+1}^s\,\mathbf{C}_k^T + \mathbf{P}_2 \end{pmatrix}.
\end{aligned} \tag{3.44}$$

Thus by Lemma A.2, the marginal distribution of $\mathbf{x}_k$ is given as

$$p(\mathbf{x}_k \mid \mathbf{y}_{1:T}) = \mathrm{N}(\mathbf{x}_k \mid \mathbf{m}_k^s, \mathbf{P}_k^s), \tag{3.45}$$

where

$$\begin{aligned}
\mathbf{m}_k^s &= \mathbf{m}_k + \mathbf{C}_k\,(\mathbf{m}_{k+1}^s - \mathbf{A}_k\,\mathbf{m}_k) \\
\mathbf{P}_k^s &= \mathbf{P}_k + \mathbf{C}_k\,(\mathbf{P}_{k+1}^s - \mathbf{P}_{k+1}^-)\,\mathbf{C}_k^T.
\end{aligned} \tag{3.46}$$

□

Example 3.3 (Kalman smoother for Gaussian random walk)**.** *The Kalman smoother for the random walk model given in Example 3.1 is given by the equations*

$$\begin{aligned}
m_{k+1}^- &= m_k \\
P_{k+1}^- &= P_k + q \\
m_k^s &= m_k + \frac{P_k}{P_{k+1}^-}(m_{k+1}^s - m_{k+1}^-) \\
P_k^s &= P_k + \left(\frac{P_k}{P_{k+1}^-}\right)^2 [P_{k+1}^s - P_{k+1}^-],
\end{aligned} \tag{3.47}$$

where m_k and P_k are the updated mean and covariance from the Kalman filter in Example 3.2.

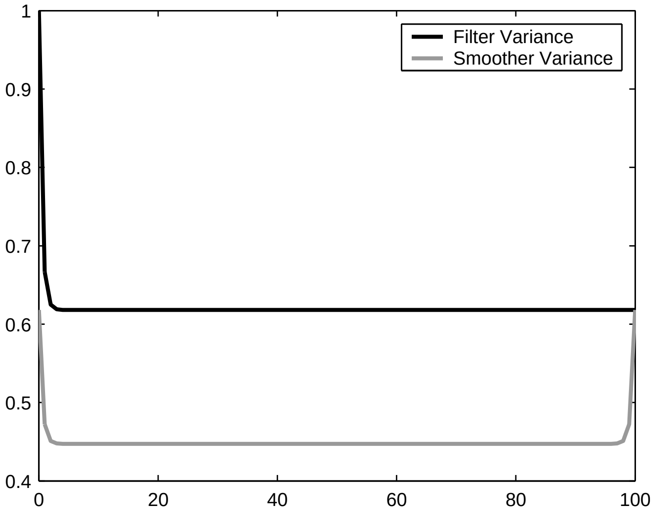

Figure 3.3: Filter and smoother variances in the Kalman smoothing example (Example 3.3).

3.1.2 Filtering and Smoothing by Gaussian Approximations

Often the dynamic and measurement processes in practical applications are not linear and the Kalman filter cannot be applied as such. However, still often the filtering and smoothing distributions of this kind of processes can be approximated with Gaussian distributions. In this section two types of methods for forming the Gaussian approximations are considered, the Taylor series based extended Kalman filters and smoothers, and unscented transform based unscented Kalman filters and smoothers.

Taylor Series Based Approximations

Next linear and quadratic approximations of transformations of Gaussian random variables are presented. These methods try to approximate the distribution of a random variable $\mathbf{y}$, which is generated as a non-linear transformation of a Gaussian random variable $\mathbf{x}$ as follows:

$$\begin{aligned} \mathbf{x} &\sim \mathrm{N}(\mathbf{m}, \mathbf{P}) \\ \mathbf{y} &= \mathbf{g}(\mathbf{x}). \end{aligned} \tag{3.48}$$

where $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{y} \in \mathbb{R}^m$, and $\mathbf{g} : \mathbb{R}^n \mapsto \mathbb{R}^m$ is a general non-linear function. Formally the probability density of the random variable $\mathbf{y}$ is[3] (see, e.g Gelman

[3] This actually only applies to invertible $\mathbf{g}(\cdot)$, but it can be easily generalized to the non-invertible case.

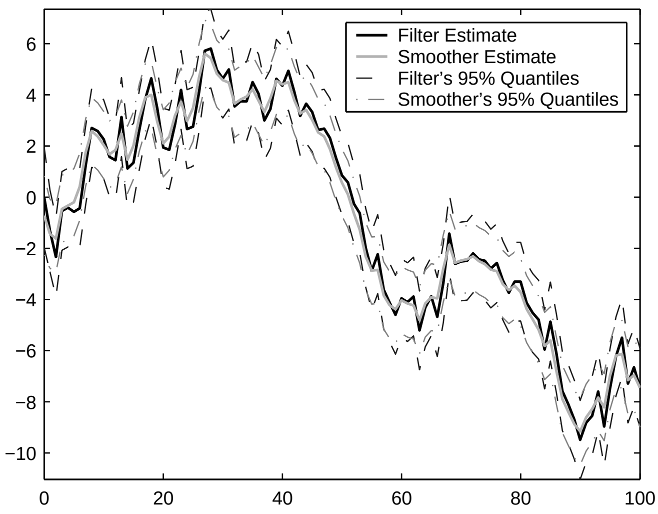

Figure 3.4: Filter and smoother estimates in the Kalman smoothing example (Example 3.3).

et al., 1995)

$$p(\mathbf{y}) = |\mathbf{J}(\mathbf{y})| \; \mathrm{N}(\mathbf{g}^{-1}(\mathbf{y}) \,|\, \mathbf{m}, \mathbf{P}), \tag{3.49}$$

where $|\mathbf{J}(\mathbf{y})|$ is the determinant of the Jacobian matrix of the inverse transform $\mathbf{g}^{-1}(\mathbf{y})$. However, it is not generally possible to handle this distribution directly, because it is non-Gaussian for all but linear $\mathbf{g}$.

Linear and quadratic approximations can be used for forming Gaussian approximations to the marginal distribution of $\mathbf{y}$ and to the joint distribution of $\mathbf{x}$ and $\mathbf{y}$. The derivations of these approximations can be found, for example, in the book of Bar-Shalom et al. (2001).

Algorithm 3.1 (Linear approximation of non-linear transform)**.** *The linear approximation based Gaussian approximation to the joint distribution of* $\mathbf{x}$ *and the transformed random variable* $\mathbf{y} = \mathbf{g}(\mathbf{x})$ *when* $\mathbf{x} \sim \mathrm{N}(\mathbf{m}, \mathbf{P})$ *is given as*

$$\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mathbf{m} \\ \boldsymbol{\mu}_L \end{pmatrix}, \begin{pmatrix} \mathbf{P} & \mathbf{C}_L \\ \mathbf{C}_L^T & \mathbf{S}_L \end{pmatrix} \right), \tag{3.50}$$

where

$$\begin{aligned} \boldsymbol{\mu}_L &= \mathbf{g}(\mathbf{m}) \\ \mathbf{S}_L &= \mathbf{G}_{\mathbf{x}}(\mathbf{m}) \, \mathbf{P} \, \mathbf{G}_{\mathbf{x}}^T(\mathbf{m}) \\ \mathbf{C}_L &= \mathbf{P} \, \mathbf{G}_{\mathbf{x}}^T(\mathbf{m}), \end{aligned} \tag{3.51}$$

and $\mathbf{G}_{\mathbf{x}}(\mathbf{m})$ *is the Jacobian matrix of* $\mathbf{g}$ *with elements*

$$\left[\mathbf{G}_{\mathbf{x}}(\mathbf{m})\right]_{j,j'} = \left.\frac{\partial g_j(\mathbf{x})}{\partial x_{j'}}\right|_{\mathbf{x}=\mathbf{m}}. \tag{3.52}$$

In quadratic approximations, in addition to the first order terms also the second order terms in the Taylor series expansion of the non-linear function are retained:

Algorithm 3.2 (Quadratic approximation of non-linear transform)**.** *The second order approximation is of the form*

$$\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix}\mathbf{m}\\ \boldsymbol{\mu}_Q\end{pmatrix}, \begin{pmatrix}\mathbf{P} & \mathbf{C}_Q\\ \mathbf{C}_Q^T & \mathbf{S}_Q\end{pmatrix}\right), \tag{3.53}$$

where the parameters are

$$\begin{aligned}
\boldsymbol{\mu}_Q &= \mathbf{g}(\mathbf{m}) + \frac{1}{2}\sum_i \mathbf{e}_i \ \mathrm{tr}\left\{\mathbf{G}_{\mathbf{xx}}^{(i)}(\mathbf{m})\,\mathbf{P}\right\} \\
\mathbf{S}_Q &= \mathbf{G}_{\mathbf{x}}(\mathbf{m})\,\mathbf{P}\,\mathbf{G}_{\mathbf{x}}^T(\mathbf{m}) + \frac{1}{2}\sum_{i,i'} \mathbf{e}_i\,\mathbf{e}_{i'}^T \ \mathrm{tr}\left\{\mathbf{G}_{\mathbf{xx}}^{(i)}(\mathbf{m})\,\mathbf{P}\,\mathbf{G}_{\mathbf{xx}}^{(i')}(\mathbf{m})\,\mathbf{P}\right\} \\
\mathbf{C}_Q &= \mathbf{P}\,\mathbf{G}_{\mathbf{x}}^T(\mathbf{m}),
\end{aligned} \tag{3.54}$$

$\mathbf{G}_{\mathbf{x}}(\mathbf{m})$ *is the Jacobian matrix* (3.52) *and* $\mathbf{G}_{\mathbf{xx}}^{(i)}(\mathbf{m})$ *is the Hessian matrix of* $g_i(\cdot)$ *evaluated at* $\mathbf{m}$*:*

$$\left[\mathbf{G}_{\mathbf{xx}}^{(i)}(\mathbf{m})\right]_{j,j'} = \left.\frac{\partial^2 g_i(\mathbf{x})}{\partial x_j\,\partial x_{j'}},\right|_{\mathbf{x}=\mathbf{m}}. \tag{3.55}$$

$\mathbf{e}_i = (0\ \cdots\ 0\ 1\ 0\ \cdots\ 0)^T$ *is a vector with 1 at position* i *and other elements are zero, that is, it is the unit vector in direction of the coordinate axis* i*.*

Discrete-Time Extended Kalman Filter

The extended Kalman filter (EKF) (see, e.g., Jazwinski, 1970; Maybeck, 1982a; Bar-Shalom et al., 2001; Grewal and Andrews, 2001) is an extension of the Kalman filter to non-linear optimal filtering problems. The filtering model is[4]

$$\begin{aligned}
\mathbf{x}_k &= \mathbf{f}(\mathbf{x}_{k-1}, k-1) + \mathbf{q}_{k-1} \\
\mathbf{y}_k &= \mathbf{h}(\mathbf{x}_k, k) + \mathbf{r}_k,
\end{aligned} \tag{3.56}$$

where $\mathbf{x}_k \in \mathbb{R}^n$ is the state, $\mathbf{y}_k \in \mathbb{R}^m$ is the measurement, $\mathbf{q}_{k-1} \sim \mathrm{N}(0, \mathbf{Q}_{k-1})$ is the Gaussian process noise, $\mathbf{r}_k \sim \mathrm{N}(0, \mathbf{R}_k)$ is the Gaussian measurement noise,

[4] In this thesis only models with additive noise are considered.

$\mathbf{f}(\cdot)$ is the dynamic model function and $\mathbf{h}(\dot{\,})$ is the measurement model function. The first and second order extended Kalman filters form Gaussian approximations

$$p(\mathbf{x}_k \,|\, \mathbf{y}_{1:k}) \approx \mathrm{N}(\mathbf{x}_k \,|\, \mathbf{m}_k, \mathbf{P}_k), \tag{3.57}$$

to the filtering densities using the linear approximation Algorithm 3.1 and quadratic approximation Algorithm 3.2, respectively.

Algorithm 3.3 (First order extended Kalman filter)**.** *The prediction and update steps of the first order extended Kalman filter are:*

- *Prediction:*

$$\begin{aligned}
\mathbf{m}_k^- &= \mathbf{f}(\mathbf{m}_{k-1}, k-1) \\
\mathbf{P}_k^- &= \mathbf{F}_\mathbf{x}(\mathbf{m}_{k-1}, k-1)\, \mathbf{P}_{k-1}\, \mathbf{F}_\mathbf{x}^T(\mathbf{m}_{k-1}, k-1) + \mathbf{Q}_{k-1}.
\end{aligned} \tag{3.58}$$

- *Update:*

$$\begin{aligned}
\mathbf{v}_k &= \mathbf{y}_k - \mathbf{h}(\mathbf{m}_k^-, k) \\
\mathbf{S}_k &= \mathbf{H}_\mathbf{x}(\mathbf{m}_k^-, k)\, \mathbf{P}_k^-\, \mathbf{H}_\mathbf{x}^T(\mathbf{m}_k^-, k) + \mathbf{R}_k \\
\mathbf{K}_k &= \mathbf{P}_k^-\, \mathbf{H}_\mathbf{x}^T(\mathbf{m}_k^-, k)\, \mathbf{S}_k^{-1} \\
\mathbf{m}_k &= \mathbf{m}_k^- + \mathbf{K}_k\, \mathbf{v}_k \\
\mathbf{P}_k &= \mathbf{P}_k^- - \mathbf{K}_k\, \mathbf{S}_k\, \mathbf{K}_k^T.
\end{aligned} \tag{3.59}$$

where the matrices $\mathbf{F}_\mathbf{x}(\mathbf{m}, k-1)$ *and* $\mathbf{H}_\mathbf{x}(\mathbf{m}, k)$ *are the Jacobian matrices of* $\mathbf{f}$ *and* $\mathbf{h}$*, respectively, with elements*

$$[\mathbf{F}_\mathbf{x}(\mathbf{m}, k-1)]_{j,j'} = \left.\frac{\partial f_j(\mathbf{x}, k-1)}{\partial x_{j'}}\right|_{\mathbf{x}=\mathbf{m}} \tag{3.60}$$

$$[\mathbf{H}_\mathbf{x}(\mathbf{m}, k)]_{j,j'} = \left.\frac{\partial h_j(\mathbf{x}, k)}{\partial x_{j'}}\right|_{\mathbf{x}=\mathbf{m}}. \tag{3.61}$$

In the second order EKF the non-linearity is approximated with quadratic approximation Algorithm 3.2.

Algorithm 3.4 (Second order extended Kalman filter)**.** *The prediction and update steps of the second order extended Kalman filter are:*

- *Prediction:*

$$
\begin{aligned}
\mathbf{m}_k^- &= \mathbf{f}(\mathbf{m}_{k-1}, k-1) + \frac{1}{2} \sum_i \mathbf{e}_i \operatorname{tr}\left\{ \mathbf{F}_{\mathbf{xx}}^{(i)}(\mathbf{m}_{k-1}, k-1)\, \mathbf{P}_{k-1} \right\} \\
\mathbf{P}_k^- &= \mathbf{F}_{\mathbf{x}}(\mathbf{m}_{k-1}, k-1)\, \mathbf{P}_{k-1}\, \mathbf{F}_{\mathbf{x}}^T(\mathbf{m}_{k-1}, k-1) \\
&\quad + \frac{1}{2} \sum_{i,i'} \mathbf{e}_i\, \mathbf{e}_{i'}^T \operatorname{tr}\left\{ \mathbf{F}_{\mathbf{xx}}^{(i)}(\mathbf{m}_{k-1}, k-1) \mathbf{P}_{k-1} \mathbf{F}_{\mathbf{xx}}^{(i')}(\mathbf{m}_{k-1}, k-1) \mathbf{P}_{k-1} \right\} \\
&\quad + \mathbf{Q}_{k-1}.
\end{aligned}
\tag{3.62}
$$

- *Update:*

$$
\begin{aligned}
\mathbf{v}_k &= \mathbf{y}_k - \mathbf{h}(\mathbf{m}_k^-, k) - \frac{1}{2} \sum_i \mathbf{e}_i \operatorname{tr}\left\{ \mathbf{H}_{\mathbf{xx}}^{(i)}(\mathbf{m}_k^-, k)\, \mathbf{P}_k^- \right\} \\
\mathbf{S}_k &= \mathbf{H}_{\mathbf{x}}(\mathbf{m}_k^-, k)\, \mathbf{P}_k^-\, \mathbf{H}_{\mathbf{x}}^T(\mathbf{m}_k^-, k) \\
&\quad + \frac{1}{2} \sum_{i,i'} \mathbf{e}_i\, \mathbf{e}_{i'}^T \operatorname{tr}\left\{ \mathbf{H}_{\mathbf{xx}}^{(i)}(\mathbf{m}_k^-, k)\, \mathbf{P}_k^-\, \mathbf{H}_{\mathbf{xx}}^{(i')}(\mathbf{m}_k^-, k)\, \mathbf{P}_k^- \right\} + \mathbf{R}_k \\
\mathbf{K}_k &= \mathbf{P}_k^-\, \mathbf{H}_{\mathbf{x}}^T(\mathbf{m}_k^-, k)\, \mathbf{S}_k^{-1} \\
\mathbf{m}_k &= \mathbf{m}_k^- + \mathbf{K}_k\, \mathbf{v}_k \\
\mathbf{P}_k &= \mathbf{P}_k^- - \mathbf{K}_k\, \mathbf{S}_k\, \mathbf{K}_k^T.
\end{aligned}
\tag{3.63}
$$

where matrices $\mathbf{F}_{\mathbf{x}}(\mathbf{m}, k-1)$ *and* $\mathbf{H}_{\mathbf{x}}(\mathbf{m}, k)$ *are given by Equations* (3.60) *and* (3.61)*. The matrices* $\mathbf{F}_{\mathbf{xx}}^{(i)}(\mathbf{m}, k-1)$ *and* $\mathbf{H}_{\mathbf{xx}}^{(i)}(\mathbf{m}, k)$ *are the Hessian matrices of* f_i *and* h_i *respectively:*

$$
\left[\mathbf{F}_{\mathbf{xx}}^{(i)}(\mathbf{m}, k-1) \right]_{j,j'} = \left. \frac{\partial^2 f_i(\mathbf{x}, k-1)}{\partial x_j\, \partial x_{j'}} \right|_{\mathbf{x}=\mathbf{m}} \tag{3.64}
$$

$$
\left[\mathbf{H}_{\mathbf{xx}}^{(i)}(\mathbf{m}, k) \right]_{j,j'} = \left. \frac{\partial^2 h_i(\mathbf{x}, k)}{\partial x_j\, \partial x_{j'}} \right|_{\mathbf{x}=\mathbf{m}}. \tag{3.65}
$$

Discrete-Time Extended Kalman Smoother

The first order (i.e., linearized) extended Kalman smoother (Cox, 1964; Sage and Melsa, 1971) can be obtained from the Kalman smoother equations by replacing the prediction equations with the first order approximations. Higher order extended Kalman smoothers are also possible (see, e.g., Cox, 1964; Sage and Melsa, 1971), but only the first order version is presented here.

Algorithm 3.5 (Extended Kalman smoother)**.** *The equations for the extended Kalman smoother are*

$$\begin{aligned}
\mathbf{m}_{k+1}^{-} &= \mathbf{f}(\mathbf{m}_k, k) \\
\mathbf{P}_{k+1}^{-} &= \mathbf{F}_{\mathbf{x}}(\mathbf{m}_k, k)\, \mathbf{P}_k\, \mathbf{F}_{\mathbf{x}}^T(\mathbf{m}_k, k) + \mathbf{Q}_k \\
\mathbf{C}_k &= \mathbf{P}_k\, \mathbf{F}_{\mathbf{x}}^T(\mathbf{m}_k, k)\, [\mathbf{P}_{k+1}^{-}]^{-1} \\
\mathbf{m}_k^s &= \mathbf{m}_k + \mathbf{C}_k\, [\mathbf{m}_{k+1}^s - \mathbf{m}_{k+1}^{-}] \\
\mathbf{P}_k^s &= \mathbf{P}_k + \mathbf{C}_k\, [\mathbf{P}_{k+1}^s - \mathbf{P}_{k+1}^{-}]\, \mathbf{C}_k^T,
\end{aligned} \tag{3.66}$$

where the matrix $\mathbf{F}_{\mathbf{x}}(\mathbf{m}_k, k)$ *is given by the Equation* (3.60).

Unscented Transform

The *unscented transform* (UT) (see, e.g., Julier and Uhlmann, 1995, 2004b; Wan and van der Merwe, 2001) can be used for forming a Gaussian approximation to the joint distribution of random variables $\mathbf{x}$ and $\mathbf{y}$, when the random variable $\mathbf{y}$ is obtained by the non-linear transformation of the Gaussian random variable $\mathbf{x}$ in the equation (3.48). The idea of UT is to form a fixed number of deterministically chosen sigma-points, which capture the mean and covariance of the original distribution of $\mathbf{x}$ exactly. These sigma-points are then propagated through the non-linearity and the mean and covariance of the transformed variable are estimated from them. Note that although the unscented transform resembles Monte Carlo estimation the approaches are significantly different, because in UT the sigma points are selected deterministically (Julier and Uhlmann, 2004b).

As discussed in Julier and Uhlmann (2004b) the unscented transform is able to capture the higher order moments caused by the non-linear transform better than the Taylor series based approximations. Other advantages of using UT instead of linearization are discussed in Julier and Uhlmann (2004b).

Algorithm 3.6 (Unscented transform)**.** *The unscented transform can be used for forming Gaussian approximation*

$$\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mathbf{m} \\ \boldsymbol{\mu}_U \end{pmatrix}, \begin{pmatrix} \mathbf{P} & \mathbf{C}_U \\ \mathbf{C}_U^T & \mathbf{S}_U \end{pmatrix} \right), \tag{3.67}$$

to the joint density of $\mathbf{x}$ *and* $\mathbf{y}$*. The unscented transform is the following:*

1. *Compute the set of* $2n{+}1$ *points from the columns of the matrix* $\sqrt{(n+\lambda)\,\mathbf{P}}$*:*

$$\begin{aligned}
\mathbf{x}^{(0)} &= \mathbf{m} \\
\mathbf{x}^{(i)} &= \mathbf{m} + \left[\sqrt{(n+\lambda)\,\mathbf{P}}\right]_i, \quad i = 1, \ldots, n \\
\mathbf{x}^{(i)} &= \mathbf{m} - \left[\sqrt{(n+\lambda)\,\mathbf{P}}\right]_i, \quad i = n+1, \ldots, 2n
\end{aligned} \tag{3.68}$$

and the associated weights:

$$\begin{aligned}
W_0^{(m)} &= \lambda/(n+\lambda) \\
W_0^{(c)} &= \lambda/(n+\lambda) + (1-\alpha^2+\beta) \\
W_i^{(m)} &= 1/\{2(n+\lambda)\}, \quad i = 1,\ldots,2n \\
W_i^{(c)} &= 1/\{2(n+\lambda)\}, \quad i = 1,\ldots,2n.
\end{aligned} \tag{3.69}$$

Parameter λ is a scaling parameter defined as

$$\lambda = \alpha^2\,(n+\kappa) - n. \tag{3.70}$$

The positive constants α, β and κ are are used as parameters of the method.

2. *Transform each of the sigma points as*

$$\mathbf{y}^{(i)} = \mathbf{g}(\mathbf{x}^{(i)}), \quad i = 0,\ldots,2n. \tag{3.71}$$

3. *Mean and covariance estimates for $\mathbf{y}$ can be calculated as*

$$\boldsymbol{\mu}_U \approx \sum_{i=0}^{2n} W_i^{(m)}\,\mathbf{y}^{(i)} \tag{3.72}$$

$$\mathbf{S}_U \approx \sum_{i=0}^{2n} W_i^{(c)}\,(\mathbf{y}^{(i)} - \boldsymbol{\mu}_U)\,(\mathbf{y}^{(i)} - \boldsymbol{\mu}_U)^T. \tag{3.73}$$

4. *The cross-covariance between $\mathbf{x}$ and $\mathbf{y}$ can estimated as*

$$\mathbf{C}_U \approx \sum_{i=0}^{2n} W_i^{(c)}\,(\mathbf{x}^{(i)} - \mathbf{m})\,(\mathbf{y}^{(i)} - \boldsymbol{\mu}_U)^T. \tag{3.74}$$

The matrix square root of positive definite matrix $\mathbf{P}$ means a matrix $\mathbf{A} = \sqrt{\mathbf{P}}$ such that

$$\mathbf{P} = \mathbf{A}\,\mathbf{A}^T. \tag{3.75}$$

Because the only requirement for $\mathbf{A}$ is the definition above, we can, for example, use the lower triangular matrix of the Cholesky factorization *(see Appendix A.2).*

Lemma 3.1 (The matrix form of UT)**.** *The unscented transform can be written in matrix form as follows:*

$$\mathbf{X} = \begin{bmatrix}\mathbf{m} & \cdots & \mathbf{m}\end{bmatrix} + \sqrt{c}\begin{bmatrix}\mathbf{0} & \sqrt{\mathbf{P}} & -\sqrt{\mathbf{P}}\end{bmatrix} \tag{3.76}$$

$$\mathbf{Y} = \mathbf{g}(\mathbf{X}) \tag{3.77}$$

$$\boldsymbol{\mu}_U = \mathbf{Y}\,\mathbf{w}_m \tag{3.78}$$

$$\mathbf{S}_U = \mathbf{Y}\,\mathbf{W}\,\mathbf{Y}^T \tag{3.79}$$

$$\mathbf{C}_U = \mathbf{X}\,\mathbf{W}\,\mathbf{Y}^T, \tag{3.80}$$

where $\mathbf{X}$ *is the matrix of sigma points, function* $\mathbf{g}(\cdot)$ *is applied to each column of the argument matrix separately,* $c = \alpha^2\,(n+\kappa)$*, and vector* $\mathbf{w}_m$ *and matrix* $\mathbf{W}$ *are defined as follows:*

$$\begin{aligned}
\mathbf{w}_m &= \begin{bmatrix} W_m^{(0)} & \cdots & W_m^{(2n)} \end{bmatrix}^T && (3.81)\\
\mathbf{W} &= \left(\mathbf{I} - \begin{bmatrix}\mathbf{w}_m & \cdots & \mathbf{w}_m\end{bmatrix}\right)\\
&\quad \times \operatorname{diag}(W_c^{(0)} \cdots W_c^{(2n)})\\
&\quad \times \left(\mathbf{I} - \begin{bmatrix}\mathbf{w}_m & \cdots & \mathbf{w}_m\end{bmatrix}\right)^T. && (3.82)
\end{aligned}$$

Proof. If we define the matrix of sigma points as

$$\mathbf{X} = \begin{bmatrix}\mathbf{x}^{(0)} & \cdots & \mathbf{x}^{(2n)}\end{bmatrix}, \tag{3.83}$$

then the sigma point computation in equations (3.68) can be written in form of Equation (3.76). The Equation (3.77) is simply the vector form of the Equation (3.71).

If we define the weight vector $\mathbf{w}_m$ and matrix $\mathbf{W}_c$ as in equations (3.81) and (3.82), respectively, and denote the matrix of sigma points of $\mathbf{y}$ as $\mathbf{Y}$ the transformed mean and covariance equations can be written as

$$\begin{aligned}
\boldsymbol{\mu}_U &= \sum_i W_m^{(i)}\,\mathbf{y}^{(i)}\\
&= \mathbf{Y}\,\mathbf{w}_m && (3.84)\\
\mathbf{S}_U &= \sum_i W_c^{(i)}\,(\mathbf{y}^{(i)} - \boldsymbol{\mu}_U)\,(\mathbf{y}^{(i)} - \boldsymbol{\mu}_U)^T\\
&= \sum_i W_c^{(i)}\,(\mathbf{y}^{(i)} - \mathbf{Y}\mathbf{w}_m)\,(\mathbf{y}^{(i)} - \mathbf{Y}\mathbf{w}_m)^T\\
&= \left(\mathbf{Y} - \mathbf{Y}\begin{bmatrix}\mathbf{w}_m & \cdots & \mathbf{w}_m\end{bmatrix}\right)\\
&\quad \times \operatorname{diag}(W_c^{(0)} \cdots W_c^{(2n)})\\
&\quad \times \left(\mathbf{Y} - \mathbf{Y}\begin{bmatrix}\mathbf{w}_m & \cdots & \mathbf{w}_m\end{bmatrix}\right)^T\\
&= \mathbf{Y}\,\mathbf{W}\,\mathbf{Y}^T && (3.85)\\
\mathbf{C}_U &= \sum_i W_c^{(i)}\,(\mathbf{x}^{(i)} - \mathbf{m})\,(\mathbf{y}^{(i)} - \boldsymbol{\mu}_U)^T\\
&= \sum_i W_c^{(i)}\,(\mathbf{x}^{(i)} - \mathbf{X}\mathbf{w}_m)\,(\mathbf{y}^{(i)} - \mathbf{Y}\mathbf{w}_m)^T\\
&= \left(\mathbf{X} - \mathbf{X}\begin{bmatrix}\mathbf{w}_m & \cdots & \mathbf{w}_m\end{bmatrix}\right)\\
&\quad \times \operatorname{diag}(W_c^{(0)} \cdots W_c^{(2n)})\\
&\quad \times \left(\mathbf{Y} - \mathbf{Y}\begin{bmatrix}\mathbf{w}_m & \cdots & \mathbf{w}_m\end{bmatrix}\right)^T\\
&= \mathbf{X}\,\mathbf{W}\,\mathbf{Y}^T, && (3.86)
\end{aligned}$$

which leads to equations (3.78), (3.79) and (3.80). □

Unscented Kalman Filter

The *unscented Kalman filter* (UKF) (Julier et al., 1995; Julier and Uhlmann, 2004b; Wan and van der Merwe, 2001) is a discrete-time optimal filtering algorithm, which utilizes the *unscented transform* for computing Gaussian approximations to the filtering solutions of non-linear optimal filtering problems of the form[5]

$$\begin{aligned} \mathbf{x}_k &= \mathbf{f}(\mathbf{x}_{k-1}, k-1) + \mathbf{q}_{k-1} && (3.87) \\ \mathbf{y}_k &= \mathbf{h}(\mathbf{x}_k, k) + \mathbf{r}_k, && (3.88) \end{aligned}$$

where $\mathbf{x}_k \in \mathbb{R}^n$ is the state, $\mathbf{y}_k \in \mathbb{R}^m$ is the measurement, $\mathbf{q}_{k-1} \sim \mathrm{N}(\mathbf{0}, \mathbf{Q}_{k-1})$ is the Gaussian process noise, and $\mathbf{r}_k \sim \mathrm{N}(\mathbf{0}, \mathbf{R}_k)$ is the Gaussian measurement noise.

Algorithm 3.7 (Unscented Kalman filter)**.** *Using the matrix form of the unscented transform (see Lemma 3.1) the UKF prediction and update steps can be written as follows:*

- Prediction: *Compute the predicted state mean* $\mathbf{m}_k^-$ *and the predicted covariance* $\mathbf{P}_k^-$ *as*

$$\begin{aligned} \mathbf{X}_{k-1} &= \begin{bmatrix} \mathbf{m}_{k-1} & \cdots & \mathbf{m}_{k-1} \end{bmatrix} + \sqrt{c} \begin{bmatrix} \mathbf{0} & \sqrt{\mathbf{P}_{k-1}} & -\sqrt{\mathbf{P}_{k-1}} \end{bmatrix} \\ \hat{\mathbf{X}}_k &= \mathbf{f}(\mathbf{X}_{k-1}, k-1) \\ \mathbf{m}_k^- &= \hat{\mathbf{X}}_k \, \mathbf{w}_m \\ \mathbf{P}_k^- &= \hat{\mathbf{X}}_k \, \mathbf{W} \, [\hat{\mathbf{X}}_k]^T + \mathbf{Q}_{k-1}. \end{aligned} \tag{3.89}$$

- Update: *Compute the predicted mean* $\boldsymbol{\mu}_k$ *and covariance of the measurement* $\mathbf{S}_k$*, and the cross-covariance of the state and measurement* $\mathbf{C}_k$*:*

$$\begin{aligned} \mathbf{X}_k^- &= \begin{bmatrix} \mathbf{m}_k^- & \cdots & \mathbf{m}_k^- \end{bmatrix} + \sqrt{c} \begin{bmatrix} \mathbf{0} & \sqrt{\mathbf{P}_k^-} & -\sqrt{\mathbf{P}_k^-} \end{bmatrix} \\ \mathbf{Y}_k^- &= \mathbf{h}(\mathbf{X}_k^-, k) \\ \boldsymbol{\mu}_k &= \mathbf{Y}_k^- \, \mathbf{w}_m \\ \mathbf{S}_k &= \mathbf{Y}_k^- \, \mathbf{W} \, [\mathbf{Y}_k^-]^T + \mathbf{R}_k \\ \mathbf{C}_k &= \mathbf{X}_k^- \, \mathbf{W} \, [\mathbf{Y}_k^-]^T. \end{aligned} \tag{3.90}$$

[5] In this thesis only the case of additive noise is considered, but UKF can also be applied to more general filtering problems with non-additive noise.

Then compute the filter gain $\mathbf{K}_k$ *and the state mean* $\mathbf{m}_k$ *and covariance* $\mathbf{P}_k$*, conditional to the measurement* $\mathbf{y}_k$ *(i.e., the updated mean and covariance):*

$$\begin{aligned}\mathbf{K}_k &= \mathbf{C}_k\,\mathbf{S}_k^{-1} \\ \mathbf{m}_k &= \mathbf{m}_k^- + \mathbf{K}_k\,[\mathbf{y}_k - \boldsymbol{\mu}_k] \\ \mathbf{P}_k &= \mathbf{P}_k^- - \mathbf{K}_k\,\mathbf{S}_k\,\mathbf{K}_k^T.\end{aligned} \tag{3.91}$$

A thorough treatment of the unscented Kalman filtering, sigma-point filtering in general and connections to several other filtering algorithms can be found in the PhD thesis by van der Merwe (2004). In the thesis, van der Merwe (2004) also presents efficient square root versions of UKF.

Unscented Kalman Smoother

The *unscented Kalman smoother* is a Gaussian approximation based smoother, where the conventional linearization of the *extended Kalman smoother* (see, e.g., Cox, 1964; Sage and Melsa, 1971) is replaced with the unscented transform. This idea of the unscented Kalman smoother is presented, for example in (Wan and van der Merwe, 2001), but without explicit equations. In the following these equations will be presented in terms of the matrix form unscented transform. These smoothing equations can be derived from the UKF prediction equations in the same way as the first order extended Kalman smoother equations and for this reason it is possible that alternative (higher order) forms of the equations could be developed.

Algorithm 3.8 (Unscented Kalman smoother)**.** *Using the matrix form of the unscented transform (see Lemma 3.1) the* unscented Kalman smoother *prediction and update steps can be written as follows:*

- Prediction: *Compute the predicted state mean* $\mathbf{m}_{k+1}^-$ *and the predicted covariance* $\mathbf{P}_{k+1}^-$*, and the cross covariance* $\mathbf{C}_{k+1}$ *as*

$$\begin{aligned}\mathbf{X}_k &= \begin{bmatrix}\mathbf{m}_k & \cdots & \mathbf{m}_k\end{bmatrix} + \sqrt{c}\begin{bmatrix}\mathbf{0} & \sqrt{\mathbf{P}_k} & -\sqrt{\mathbf{P}_k}\end{bmatrix} \\ \hat{\mathbf{X}}_{k+1} &= \mathbf{f}(\mathbf{X}_k, k) \\ \mathbf{m}_{k+1}^- &= \hat{\mathbf{X}}_{k+1}\,\mathbf{w}_m \\ \mathbf{P}_{k+1}^- &= \hat{\mathbf{X}}_{k+1}\,\mathbf{W}\,[\hat{\mathbf{X}}_{k+1}]^T + \mathbf{Q}_{k-1} \\ \mathbf{C}_{k+1} &= \mathbf{X}_k\,\mathbf{W}\,[\hat{\mathbf{X}}_{k+1}]^T.\end{aligned} \tag{3.92}$$

where $\mathbf{m}_k$ *and* $\mathbf{P}_k$ *are the mean and covariance estimates computed by the unscented Kalman filter. Note that this prediction is the same as the unscented Kalman filter prediction step and thus we can also store the predicted means and covariances and cross-covariances in the filter in addition to the updated means and covariances.*

- Update: *Compute the smoother gain* $\mathbf{D}_k$*, and the smoothed mean* $\mathbf{m}_k^s$ *and the covariance* $\mathbf{P}_k^s$*:*

$$\begin{aligned}\mathbf{D}_k &= \mathbf{C}_{k+1}\,[\mathbf{P}_{k+1}^-]^{-1} \\ \mathbf{m}_k^s &= \mathbf{m}_k + \mathbf{D}_k\,\left[\mathbf{m}_{k+1}^s - \mathbf{m}_{k+1}^-\right] \\ \mathbf{P}_k^s &= \mathbf{P}_k + \mathbf{D}_k\,\left[\mathbf{P}_{k+1}^s - \mathbf{P}_{k+1}^-\right]\,\mathbf{D}_k^T.\end{aligned} \tag{3.93}$$

3.1.3 Sequential Importance Resampling

Although in many problems Gaussian distributions approximate well the filtering and smoothing distributions, sometimes these distributions can be, for example, multi-modal in which case Gaussian approximations do not work well. In this kind of cases and, for example, when some of the state components are discrete, sequential importance resampling based particle filter and smoother approximations can be a better alternative. This section considers particle filters, which are methods for forming Monte Carlo approximations to the optimal filtering and smoothing solutions.

Sequential Importance Resampling Filter

Sequential importance resampling (SIR)[6] (Gordon et al., 1993; Kitagawa, 1996; Doucet et al., 2001; Ristic et al., 2004), is a generalization of the *particle filtering* framework for the estimation of generic state space models of the form

$$\begin{aligned}\mathbf{x}_k &\sim p(\mathbf{x}_k \mid \mathbf{x}_{k-1}) \\ \mathbf{y}_k &\sim p(\mathbf{y}_k \mid \mathbf{x}_k),\end{aligned} \tag{3.94}$$

where $\mathbf{x}_k \in \mathbb{R}^n$ is the state on time step k and $\mathbf{y}_k \in \mathbb{R}^m$ is the measurement. The state and measurements may contain both discrete and continuous components.

The SIR algorithm uses a weighted set of particles $\{(w_k^{(i)}, \mathbf{x}_k^{(i)}) \;:\; i = 1, \ldots, N\}$ for representing the filtering distribution $p(\mathbf{x}_k \,|\, \mathbf{y}_{1:k})$ such that on every time step k an approximation of the expectation of an arbitrary function $\mathbf{g}(\mathbf{x})$ can be calculated as the weighted sample average

$$E[\mathbf{g}(\mathbf{x}_k) \,|\, \mathbf{y}_{1:k}] \approx \sum_{i=1}^N w_k^{(i)} \mathbf{g}(\mathbf{x}_k^{(i)}). \tag{3.95}$$

Equivalently, SIR can be interpreted to form an approximation of the posterior distribution as

$$p(\mathbf{x}_k \,|\, \mathbf{y}_{1:k}) \approx \sum_{i=1}^N w_k^{(i)} \delta(\mathbf{x}_k - \mathbf{x}_k^{(i)}), \tag{3.96}$$

[6] *Sequential importance resampling (SIR)* is also often referred to as *sampling importance resampling (SIR) or* sequential importance sampling resampling (SISR).

where $\delta(\cdot)$ is the Dirac delta function.

The set of particles is updated and reweighted using a recursive version of importance sampling. An additional *resampling* procedure is used for removing particles with very small weights and duplicating particles with large weights. The variance introduced by the resampling procedure can be reduced by proper choice of the resampling method. The *stratified resampling* algorithm (Kitagawa, 1996) is optimal in terms of variance.

Usually resampling is not performed on every time step, but only when it is actually needed. One way of implementing this is to do resampling on every nth step, where n is some predefined constant. This method has the advantage that it is unbiased. Another way, which is used in this thesis, is the *adaptive resampling*. In this method the effective number of particles, which is estimated from the variance of the particle weights (Liu and Chen, 1995), is used for monitoring the need for resampling. An estimate for the effective number of particles based on an approximation of the variance of importance weights can be computed as:

$$n_{\text{eff}} \approx \frac{1}{\sum_{i=1}^{N} \left(w_k^{(i)} \right)^2}, \tag{3.97}$$

where $w_k^{(i)}$ is the normalized weight of particle i on the time step k (Liu and Chen, 1995). Resampling is performed when the effective number of particles is significantly less than the total number of particles, for example, $n_{\text{eff}} < N/10$, where N is the total number of particles.

Algorithm 3.9 (Sequential importance resampling)**.** *The SIR algorithm can be summarized as follows:*

1. *Draw new point* $\mathbf{x}_k^{(i)}$ *for each point in the sample set* $\{\mathbf{x}_{k-1}^{(i)}, i = 1, \ldots, N\}$ *from the importance distribution:*

$$\mathbf{x}_k^{(i)} \sim \pi(\mathbf{x}_k \mid \mathbf{x}_{k-1}^{(i)}, \mathbf{y}_{1:k}), \qquad i = 1, \ldots, N. \tag{3.98}$$

2. *Calculate new weights*

$$w_k^{(i)} \propto w_{k-1}^{(i)} \frac{p(\mathbf{y}_k \mid \mathbf{x}_k^{(i)})\, p(\mathbf{x}_k^{(i)} \mid \mathbf{x}_{k-1}^{(i)})}{\pi(\mathbf{x}_k^{(i)} \mid \mathbf{x}_{k-1}^{(i)}, \mathbf{y}_{1:k})}, \qquad i = 1, \ldots, N. \tag{3.99}$$

   and normalize them to sum to unity.

3. *If the effective number of particles* (3.97) *is too low, perform resampling.*

The performance of the SIR algorithm is dependent on the importance distribution $\pi(\cdot)$, which is an approximation of posterior distribution of states given the values at the previous step. The importance distribution should be in such functional form that it is easy to draw samples from it and it is possible to evaluate the probability densities of the sample points. *The optimal importance distribution* in terms of variance (see, e.g., Doucet et al., 2001; Ristic et al., 2004) is

$$\pi(\mathbf{x}_k \mid \mathbf{x}_{k-1}, \mathbf{y}_{1:k}) = p(\mathbf{x}_k \mid \mathbf{x}_{k-1}, \mathbf{y}_{1:k}). \tag{3.100}$$

If the optimal importance distribution cannot be directly used, good importance distributions can be obtained by *local linearization* where a mixture of extended Kalman filters (EKF) or unscented Kalman filters (UKF) is used as the importance distribution (Doucet et al., 2000; van der Merwe et al., 2001). van der Merwe et al. (2001) also suggest a Metropolis-Hastings step after (or in place of) resampling step to smooth the resulting distribution, but from their results, it seems that this extra computation step has no significant performance effect. A particle filter with UKF importance distribution is also referred to as *unscented particle filter* (UPF).

The bootstrap filter (Gordon et al., 1993) is a variation of SIR, where the dynamic model $p(\mathbf{x}_k \mid \mathbf{x}_{k-1})$ is used as the importance distribution. This makes the implementation of the algorithm very easy, but due to the inefficiency of the importance distribution it may require a very large number of Monte Carlo samples for accurate estimation results.

By tuning the resampling algorithm to specific estimation problems and possibly changing the order of weight computation and sampling, accuracy and computational efficiency of the algorithm can be improved (Fearnhead and Clifford, 2003). An important issue is that sampling is more efficient without replacement, such that duplicate samples are not stored. There is also evidence that in some situations it is more efficient to use a simple deterministic algorithm for preserving the N most likely particles. In the article (Punskaya et al., 2002) it is shown that in digital demodulation, where the sampled space is discrete and the optimization criterion is the minimum error, the deterministic algorithm performs better.

Sequential Importance Resampling Smoother

Optimal smoothing can be performed with the SIR algorithm with a slight modification to the filtering case. Instead of keeping Monte Carlo samples of the states on single time step $\mathbf{x}_k^{(i)}$, we keep samples of the whole state histories $\mathbf{x}_{1:k}^{(i)}$. The computations of the algorithm remain exactly the same, but in resampling stage the whole state histories are resampled instead of the states of single time steps. The weights of these state histories are the same as in normal SIR algorithm and the smoothed posterior distribution estimate of time step k given the measurements up to the time step $T > k$ is given as (Kitagawa, 1996; Doucet et al., 2000)

$$p(\mathbf{x}_k \,|\, \mathbf{y}_{1:T}) \approx \sum_{i=1}^{N} w_T^{(i)} \delta(\mathbf{x}_k - \mathbf{x}_k^{(i)}). \tag{3.101}$$

where $\delta(\cdot)$ is the Dirac delta function and $\mathbf{x}_k^{(i)}$ is the kth component in $\mathbf{x}_{1:T}^{(i)}$.

However, if $T \gg k$ this simple method is known to produce very degenerate approximations (Kitagawa, 1996; Doucet et al., 2000). In (Godsill et al., 2004) more efficient methods for sampling from the smoothing distributions are presented.

Rao-Blackwellized Particle Filter

One way of improving the efficiency of SIR is to use Rao-Blackwellization. The idea of the *Rao-Blackwellized particle filter* (RBPF) (Akashi and Kumamoto, 1977; Doucet et al., 2001; Ristic et al., 2004) is that sometimes it is possible to evaluate some of the filtering equations analytically and the others with Monte Carlo sampling instead of computing everything with pure sampling. According to the *Rao-Blackwell theorem* (see, e.g., Berger, 1985; Casella and Robert, 1996) this leads to estimators with less variance than what could be obtained with pure Monte Carlo sampling. An intuitive way of understanding this is that the marginalization replaces the finite Monte Carlo particle set representation with an infinite closed form particle set, which is always more accurate than any finite set.

Most commonly Rao-Blackwellized particle filtering refers to marginalized filtering of conditionally Gaussian Markov models of the form

$$\begin{aligned} p(\mathbf{x}_k \,|\, \mathbf{x}_{k-1}, \boldsymbol{\theta}_{k-1}) &= \mathrm{N}(\mathbf{x}_k \,|\, \mathbf{A}_{k-1}(\boldsymbol{\theta}_{k-1})\, \mathbf{x}_{k-1}, \mathbf{Q}_{k-1}(\boldsymbol{\theta}_{k-1})) \\ p(\mathbf{y}_k \,|\, \mathbf{x}_k, \boldsymbol{\theta}_k) &= \mathrm{N}(\mathbf{y}_k \,|\, \mathbf{H}_k(\boldsymbol{\theta}_k)\, \mathbf{x}_k, \mathbf{R}_k(\boldsymbol{\theta}_k)) \\ p(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}_{k-1}) &= \text{(any given form)}, \end{aligned} \tag{3.102}$$

where $\mathbf{x}_k$ is the state, $\mathbf{y}_k$ is the measurement, and $\boldsymbol{\theta}_k$ is an arbitrary latent variable. If also the prior of $\mathbf{x}_k$ is Gaussian, due to conditionally Gaussian structure of the model the state variables $\mathbf{x}_k$ can be integrated out analytically and only the latent variables $\boldsymbol{\theta}_k$ need to be sampled. The Rao-Blackwellized particle filter uses SIR for the latent variables and computes everything else in closed form.

Algorithm 3.10 (Conditionally Gaussian Rao-Blackwellized particle filter)**.** *Given an importance distribution $\pi(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}_{1:k-1}^{(i)}, \mathbf{y}_{1:k})$ and a set of weighted samples $\{w_{k-1}^{(i)}, \boldsymbol{\theta}_{k-1}^{(i)}, \mathbf{m}_{k-1}^{(i)}, \mathbf{P}_{k-1}^{(i)} \;:\; i = 1, \ldots, N\}$, the Rao-Blackwellized particle filter processes each measurement $\mathbf{y}_k$ as follows (Doucet et al., 2001):*

1. *Perform Kalman filter predictions for each of the Kalman filter means and covariances in the particles $i = 1, \ldots, N$ conditional on the previously drawn latent variable values $\boldsymbol{\theta}_{k-1}^{(i)}$*

$$\begin{aligned} \mathbf{m}_k^{-(i)} &= \mathbf{A}_{k-1}(\boldsymbol{\theta}_{k-1}^{(i)})\, \mathbf{m}_{k-1}^{(i)} \\ \mathbf{P}_k^{-(i)} &= \mathbf{A}_{k-1}(\boldsymbol{\theta}_{k-1}^{(i)})\, \mathbf{P}_{k-1}^{(i)}\, \mathbf{A}_{k-1}^T(\boldsymbol{\theta}_{k-1}^{(i)}) + \mathbf{Q}_{k-1}(\boldsymbol{\theta}_{k-1}^{(i)}). \end{aligned} \tag{3.103}$$

2. *Draw new latent variables* $\boldsymbol{\theta}_k^{(i)}$ *for each particle in* $i = 1, \ldots, N$ *from the corresponding importance distributions*

$$\boldsymbol{\theta}_k^{(i)} \sim \pi(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}_{1:k-1}^{(i)}, \mathbf{y}_{1:k}). \tag{3.104}$$

3. *Calculate new weights as follows:*

$$w_k^{(i)} \propto w_{k-1}^{(i)} \frac{p(\mathbf{y}_k \mid \boldsymbol{\theta}_{1:k}^{(i)}, \mathbf{y}_{1:k-1})\, p(\boldsymbol{\theta}_k^{(i)} \mid \boldsymbol{\theta}_{k-1}^{(i)})}{\pi(\boldsymbol{\theta}_k^{(i)} \mid \boldsymbol{\theta}_{1:k-1}^{(i)}, \mathbf{y}_{1:k})}, \tag{3.105}$$

where the likelihood term is the marginal measurement likelihood of the Kalman filter

$$\begin{aligned} &p(\mathbf{y}_k \mid \boldsymbol{\theta}_{1:k}^{(i)}, \mathbf{y}_{1:k-1}) \\ &\quad = \mathrm{N}\left(\mathbf{y}_k \,\middle|\, \mathbf{H}_k(\boldsymbol{\theta}_k^{(i)})\, \mathbf{m}_k^{-(i)}, \mathbf{H}_k(\boldsymbol{\theta}_k^{(i)})\, \mathbf{P}_k^{-(i)}\, \mathbf{H}_k^T(\boldsymbol{\theta}_k^{(i)}) + \mathbf{R}_k(\boldsymbol{\theta}_k^{(i)})\right). \end{aligned} \tag{3.106}$$

such that the model parameters in the Kalman filter are conditioned on the drawn latent variable value $\boldsymbol{\theta}_k^{(i)}$. *Then normalize the weights to sum to unity.*

4. *Perform Kalman filter updates for each of the particles conditional on the drawn latent variables* $\boldsymbol{\theta}_k^{(i)}$

$$\begin{aligned} \mathbf{v}_k^{(i)} &= \mathbf{y}_k - \mathbf{H}_k(\boldsymbol{\theta}_k^{(i)})\, \mathbf{m}_k^{-} \\ \mathbf{S}_k^{(i)} &= \mathbf{H}_k(\boldsymbol{\theta}_k^{(i)})\, \mathbf{P}_k^{-(i)}\, \mathbf{H}_k^T(\boldsymbol{\theta}_k^{(i)}) + \mathbf{R}_k(\boldsymbol{\theta}_k^{(i)}) \\ \mathbf{K}_k^{(i)} &= \mathbf{P}_k^{-(i)}\, \mathbf{H}_k^T(\boldsymbol{\theta}_k^{(i)})\, \mathbf{S}_k^{-1} \\ \mathbf{m}_k^{(i)} &= \mathbf{m}_k^{-(i)} + \mathbf{K}_k^{(i)}\, \mathbf{v}_k^{(i)} \\ \mathbf{P}_k^{(i)} &= \mathbf{P}_k^{-(i)} - \mathbf{K}_k^{(i)}\, \mathbf{S}_k^{(i)}\, [\mathbf{K}_k^{(i)}]^T. \end{aligned} \tag{3.107}$$

5. *If the effective number of particles* (3.97) *is too low, perform* resampling.

The Rao-Blackwellized particle filter produces for each time step k a set of weighted samples $\{w_k^{(i)}, \boldsymbol{\theta}_k^{(i)}, \mathbf{m}_k^{(i)}, \mathbf{P}_k^{(i)} \;:\; i = 1, \ldots, N\}$ such that expectation of a function $\mathbf{g}(\cdot)$ can be approximated as

$$\mathrm{E}[\mathbf{g}(\mathbf{x}_k, \boldsymbol{\theta}_k) \mid \mathbf{y}_{1:k}] \approx \sum_{i=1}^{N} w_k^{(i)} \int \mathbf{g}(\mathbf{x}_k, \boldsymbol{\theta}_k^{(i)})\; \mathrm{N}(\mathbf{x}_k \mid \mathbf{m}_k^{(i)}, \mathbf{P}_k^{(i)})\, \mathrm{d}\mathbf{x}_k. \tag{3.108}$$

Equivalently the RBPF can be interpreted to form an approximation of the filtering distribution as

$$p(\mathbf{x}_k, \boldsymbol{\theta}_k \,|\, \mathbf{y}_{1:k}) \approx \sum_{i=1}^{N} w_k^{(i)} \, \delta(\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^{(i)}) \; \mathrm{N}(\mathbf{x}_k \,|\, \mathbf{m}_k^{(i)}, \mathbf{P}_k^{(i)}). \tag{3.109}$$

In some cases, when the filtering model is not strictly Gaussian due to slight non-linearities in either dynamic or measurement models it is possible to replace the exact Kalman filter update and prediction steps in RBPF with extended Kalman filter (EKF) or unscented Kalman filter (UKF) prediction and update steps.

In addition to the conditional Gaussian models, another general class of models where Rao-Blackwellization can often be applied are state space models with unknown static parameters. These models are of the form (Storvik, 2002)

$$\begin{aligned} \mathbf{x}_k &\sim p(\mathbf{x}_k \,|\, \mathbf{x}_{k-1}, \boldsymbol{\theta}) \\ \mathbf{y}_k &\sim p(\mathbf{y}_k \,|\, \mathbf{x}_k, \boldsymbol{\theta}) \\ \boldsymbol{\theta} &\sim p(\boldsymbol{\theta}), \end{aligned} \tag{3.110}$$

where vector $\boldsymbol{\theta}$ contains the unknown static parameters. If the posterior distribution of parameters $\boldsymbol{\theta}$ depends only on some sufficient statistics

$$\mathbf{T}_k = \mathbf{T}_k(\mathbf{x}_{1:k}, \mathbf{y}_{1:k}), \tag{3.111}$$

and if the sufficient statics are easy to update recursively, then sampling of the state and parameters can be efficiently performed by recursively computing the sufficient statistics conditionally to the sampled states and the measurements (Storvik, 2002).

A particularly useful special case is obtained when the dynamic model is independent of the parameters $\boldsymbol{\theta}$. In this case, if conditionally to the state $\mathbf{x}_k$ the prior $p(\boldsymbol{\theta})$ belongs to the conjugate family of the likelihood $p(\mathbf{y}_k \,|\, \mathbf{x}_k, \boldsymbol{\theta})$, the static parameters $\boldsymbol{\theta}$ can be marginalized out and only the states need to be sampled.

Rao-Blackwellized Particle Smoother

The Rao-Blackwellized particle smoother can be used for computing the smoothing solution to the conditionally Gaussian RBPF model (3.102). A weighted set of Monte Carlo samples from the smoothed distribution of the parameters $\boldsymbol{\theta}_k$ in the model (3.102) can be produced by storing the histories instead of the single states, as in the case of plain SIR. The corresponding histories of the means and the covariances are then conditional on the *parameter histories* $\boldsymbol{\theta}_{1:T}$. However, the means and covariances at time step k are only conditional on the *measurement histories* up to k, not on the later measurements. In order to correct this, Kalman smoothers have to be applied to each history of the means and the covariances.

Algorithm 3.11 (Rao-Blackwellized particle smoother)**.** *A set of weighted samples* $\{w_T^{s,(i)}, \boldsymbol{\theta}_{1:T}^{s,(i)}, \mathbf{m}_{1:T}^{s,(i)}, \mathbf{P}_{1:T}^{s,(i)} \; : \; i = 1, \ldots, N\}$ *representing the smoothed distribution can be computed as follows:*

1. *Compute the weighted set of Rao-Blackwellized state histories*

$$\{w_T^{(i)}, \boldsymbol{\theta}_{1:T}^{(i)}, \mathbf{m}_{1:T}^{(i)}, \mathbf{P}_{1:T}^{(i)} \; : \; i = 1, \ldots, N\} \tag{3.112}$$

   by using the Rao-Blackwellized particle filter.

2. *Set*

$$\begin{aligned} w_T^{s,(i)} &= w_T^{(i)} \\ \boldsymbol{\theta}_{1:T}^{s,(i)} &= \boldsymbol{\theta}_{1:T}^{(i)}. \end{aligned} \tag{3.113}$$

3. *Apply the Kalman smoother to each of the mean and covariance histories* $\mathbf{m}_{1:T}^{(i)}, \mathbf{P}_{1:T}^{(i)}$ *for* $i = 1, \ldots, N$ *to produce the smoothed mean and covariance histories* $\mathbf{m}_{1:T}^{s,(i)}, \mathbf{P}_{1:T}^{s,(i)}$.

The Rao-Blackwellized particle smoother in this simple form also has the same disadvantage as the plain SIR smoother, that is, the smoothed estimate of $\boldsymbol{\theta}_k$ can be quite degenerate if $T \gg k$. Fortunately, the smoothed estimates of the actual states $\mathbf{x}_k$ can still be quite good, because its degeneracy is avoided by the Rao-Blackwellization. To avoid the degeneracy in estimates of $\boldsymbol{\theta}_k$ it is possible to use more efficient sampling procedures for generating samples from the smoothing distributions (Fong et al., 2002).

As in the case of filtering, in some cases approximately Gaussian parts of a state space model can be approximately marginalized by using extended Kalman smoothers or unscented Kalman smoothers.

In the case of Rao-Blackwellization of static parameters (Storvik, 2002) the smoothing is much easier. In this case, due to lack of dynamics, the posterior distribution obtained after processing the last measurement is the smoothed distribution.

3.1.4 Illustrative Examples

Tracking a Sine Signal in Clutter

In this example scenario the true signal is the sine signal

$$x(t) = \sin(a\, t), \tag{3.114}$$

where the angular velocity a is only approximately known. Half of the measurements are corrupted by additive Gaussian noise with known standard deviation

$1/5$ and half of them are completely corrupted so that they can take any value in the sensor's dynamic range, which in this case is $[-2, 2]$.

Assuming the sampling period Δt, the true signal (3.114) can be approximately modeled by the discretized Wiener velocity model (see Example 2.3 on page 45)

$$\begin{pmatrix} x_k \\ \dot{x}_k \end{pmatrix} = \begin{pmatrix} 1 & \Delta t \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_{k-1} \\ \dot{x}_{k-1} \end{pmatrix} + \mathbf{q}_{k-1}.$$

where the discrete Gaussian white noise process $\mathbf{q}_{k-1}$ has the moments

$$\begin{aligned} \mathrm{E}[\mathbf{q}_{k-1}] &= \mathbf{0} \\ \mathrm{E}[\mathbf{q}_{k-1} \mathbf{q}_{k-1}^T] &= \begin{pmatrix} \frac{1}{3}\Delta t^3 & \frac{1}{2}\Delta t^2 \\ \frac{1}{2}\Delta t^2 & \Delta t \end{pmatrix} q, \end{aligned}$$

where q is the spectral density of the continuous-time white noise. The state is two dimensional $\mathbf{x}_k = (x_k \; \dot{x}_k)^T$ where x_k is the value of signal at time step $t_k = t_0 + k\Delta t$ and $\dot{x}_k$ is the derivative of signal at the same time step.

The likelihood of the measurement y_k can be modeled by defining a latent variable or *data association indicator* c_k, which has the value of 0 if the measurement is a corrupted measurement (clutter/outlier) and 1 if it is a measurement from the signal. If the measurement is clutter, it is assumed to be evenly distributed in the measurement space $[-2, 2]$ (which is the dynamic range of sensor). The likelihood is

$$p(y_k \mid \mathbf{x}_k, c_k) = \begin{cases} 1/4 & , \text{ if } c_k = 0 \\ \mathrm{N}(y_k \mid \mathbf{H}\,\mathbf{x}_k, 1/5^2) & , \text{ if } c_k = 1, \end{cases}$$

where $\mathbf{H} = \begin{pmatrix} 1 & 0 \end{pmatrix}$. The prior distributions of the signal and its derivative were chosen to be $x_0 \sim \mathrm{N}(0, 1/10^2)$ and $\dot{x}_0 \sim \mathrm{N}(1, 1/10^2)$.

This model is conditionally Gaussian given the data association indicators and thus the Rao-Blackwellized particle filter (Algorithm 3.10) can be applied. The idea of using Rao-Blackwellized particle filtering in this kind of data association problems is generalized in Section 4.1 to general multiple target tracking of unknown number of targets. This algorithm framework is here referred to as the *Rao-Blackwellized Monte Carlo data association* (RBMCDA) algorithm.

Table 3.1 shows the RMSE results of tracking the simulated sine signals with the following methods:

- *RBMCDA, 10 particles*: Rao-Blackwellized Monte Carlo data association algorithm with 10 particles.

- *RBMCDA, 100 particles*: Rao-Blackwellized Monte Carlo data association algorithm with 100 particles.

Table 3.1: Root mean squared error values for the different methods for tracking a sine signal in 50% clutter. The means (RMSE) and standard deviations (STD) from 10 different simulated data sets are given in the table. The same data sets were used with all the methods.

Method	*RMSE*	*STD*
RBMCDA, 10 particles	0.16	0.02
RBMCDA, 100 particles	0.15	0.01
Bootstrap filter, 1000 particles	2.07	2.31
Bootstrap filter, 10000 particles	0.16	0.02
Kalman filter, assuming no clutter	0.39	0.02
Kalman filter, clutter modeled	0.32	0.03
Kalman filter, perfect associations	0.11	0.01

- *Bootstrap filter, 1000 particles*: Bootstrap filter with adaptive resampling and 1000 particles, such that the joint distribution of states and data associations is represented as a set of weighted Monte Carlo samples. The high RMSE values are due to filter divergence in many of the test cases.

- *Bootstrap filter, 10000 particles*: The same bootstrap filter as above with 10000 particles.

- *Kalman filter, assuming no clutter*: Kalman filter with the assumption that there are no clutter measurements at all.

- *Kalman filter, clutter modeled*: Kalman filter with increased measurement variance such that the presence of 50% clutter is taken into account.

- *Kalman filter, perfect associations*: Kalman filter with perfect data association knowledge, such that clutter measurements are simply thrown away as would an ideal data association algorithm do.

Typical conditional means of the estimated marginal state distributions when the Rao-Blackwellized Monte Carlo data association method is used are shown in Figure 3.5. It can be seen that the estimate follows the true signal trajectory quite well despite the high number of clutter measurements.

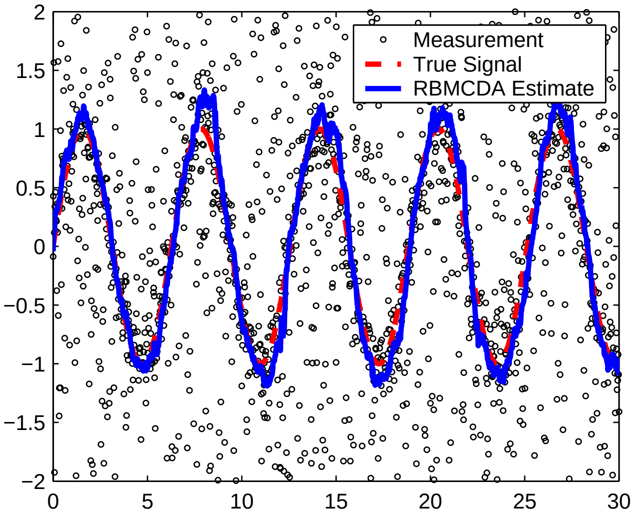

Figure 3.5: Result of tracking a sine signal in the presence of 50% clutter measurements with RBMCDA and 100 particles.

3.2 Continuous-Discrete Filtering and Smoothing

This section starts by formulating the continuous-discrete filtering as sequential solving of the Kolmogorov forward equation and the application of the Bayes' rule. The general equations of probabilistic continuous-discrete smoothing are also presented. Then the classical methods of continuous-discrete filtering are reviewed, namely, the continuous-discrete Kalman filter and the continuous-discrete extended Kalman filter. It is also shown how the more recently developed unscented Kalman filter can be applied to non-linear continuous-discrete filtering and smoothing problems. The novel continuous-discrete unscented Kalman filter and smoother are based on the new matrix form of the unscented transform. Novel algorithms for continuous-discrete sequential importance sampling are also presented, which are based on application of the Girsanov theorem.

Most of the continuous-discrete filtering problems considered in this section have the same form as in the classic book of Jazwinski (1970). The dynamics of the processes are modeled as Itô *stochastic differential equations* (SDE) driven by Brownian motions and the measurements are modeled as non-linear functions of the state, which are corrupted by Gaussian measurement noises. The continuous-discrete particle filters have the same kind of dynamic models, but the measurement model is allowed to be an arbitrary probability distribution. A bit more general class of dynamic models is implicitly covered in context of multiple target tracking in Section 4.1. However, in that section the dynamic models, which cannot be presented as Brownian motion driven stochastic differential equations are directly modeled in terms of transition probabilities, not directly as continuous-time processes.

3.2.1 Continuous-Discrete Filtering and Smoothing Equations

Next the general formal solutions to the Bayesian continuous-discrete filtering and smoothing problems are presented. Then the closed form solutions to the linear Gaussian problem, the continuous-discrete Kalman filter and Kalman smoother, are presented.

Continuous-Discrete State Space Models

Definition 3.2 (Continuous-discrete state space model)**.** *A continuous-discrete state space model is a model of the form*

$$\begin{aligned} \mathrm{d}\mathbf{x} &= \mathbf{f}(\mathbf{x}, t)\,\mathrm{d}t + \mathbf{L}(t)\,\mathrm{d}\boldsymbol{\beta}(t) \\ \mathbf{y}_k &\sim p(\mathbf{y}_k \,|\, \mathbf{x}(t_k)), \end{aligned} \tag{3.115}$$

where

- $\mathbf{x}(t) \in \mathbb{R}^n$ *is the state,*

- $\mathbf{y}_k \in \mathbb{R}^m$ *is the measurement obtained at time instance* t_k.
- $\mathbf{f} : \mathbb{R}^n \times \mathbb{R}_+ \mapsto \mathbb{R}^n$ *is the drift function,*
- $\mathbf{L}(t) \in \mathbb{R}^{n \times s}$ *is the dispersion matrix,*
- $\boldsymbol{\beta}(t) \in \mathbb{R}^s$ *is Brownian motion with diffusion matrix,* $\mathbf{Q}_c(t) \in \mathbb{R}^{s \times s}$,
- $p(\mathbf{y}_k \,|\, \mathbf{x}(t_k))$ *is the measurement model, which defines the distribution (or likelihood) of measurement* $\mathbf{y}_k$ *given the state* $\mathbf{x}(t_k)$,

Note that the dispersion term $\mathbf{L}(\cdot)$ is not allowed to depend on the state $\mathbf{x}(t)$. This is clearly a restriction, because for example in financial applications the dispersion matrix typically depends on the state. However, in the models found in navigation, tracking, control, communications and physical applications, which are the main targets of this thesis the dispersion matrices are typically independent of the state. Still the results presented here could be generalized to models with state dependent dispersion matrices.

In estimation and stochastic control context (Gelb, 1974; Maybeck, 1979, 1982a; Bar-Shalom et al., 2001; Grewal and Andrews, 2001) the stochastic differential equation in (3.115) is often stated in terms of a white noise process $\mathbf{w}(t)$ as

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{f}(\mathbf{x}, t) + \mathbf{L}(t)\,\mathbf{w}(t), \tag{3.116}$$

where the white noise is defined as the formal derivative of the Brownian motion $\mathbf{w}(t) = \mathrm{d}\boldsymbol{\beta}/\mathrm{d}t$. As already discussed in Section 2.2, the theoretical problem in this white noise formulation is that the white noise as a stochastic process cannot exists in the mathematical sense, because Brownian motion is nowhere differentiable. For this reason the integral equation formulation of the SDE as in (3.115) is often used in mathematical analysis. In practice, models are much easier to formulate in terms of white noise and for this reason it is often used in engineering and physics applications. Fortunately, all sensible models involving white noise can also be interpreted in terms of Brownian motion.

Continuous-Discrete Filtering Equations

A conceptually simple way of dealing with the continuous-discrete time filtering model is to solve the transition densities $p(\mathbf{x}(t_k) \mid \mathbf{x}(t_{k-1}))$ from the *Kolmogorov forward partial differential equation* (see, Section 2.2, Theorem 2.2, p. 38). Given the transition density $p(\mathbf{x}(t_k) \mid \mathbf{x}(t_{k-1}))$ and the measurement model $p(\mathbf{y}_k \mid \mathbf{x}(t_k))$ the discrete-time filtering equations can be applied as such. Thus the general continuous-discrete filtering algorithm can be written as follows:

Algorithm 3.12 (Continuous-discrete Bayesian filtering equations I)**.** *For each measurement* $\mathbf{y}_k$*, do the following:*

1. Discretization: *Solve the transition density* $p(\mathbf{x}(t_k) \mid \mathbf{x}(t_{k-1}))$ *from the boundary value problem*

$$\begin{aligned} \frac{\partial p}{\partial t} &= \mathcal{A}_t^*[p] \\ p(\mathbf{x}, t_{k-1}) &= \delta(\mathbf{x} - \mathbf{x}(t_{k-1})), \end{aligned} \tag{3.117}$$

where the operator $\mathcal{A}_t^*$ *is the Kolmogorov forward operator defined in Equation* (2.54) *and the transition density is interpreted as function of two variables* $p(\mathbf{x}(t) \mid \mathbf{x}(t_{k-1})) \triangleq p(\mathbf{x}, t)$ *and the final solution is the function at time* $t = t_k$*.*

2. Prediction: *Compute the predicted probability density by the Chapman-Kolmogorov equation*

$$p(\mathbf{x}(t_k) \,|\, \mathbf{y}_{1:k-1}) = \int p(\mathbf{x}(t_k) \,|\, \mathbf{x}(t_{k-1}))\, p(\mathbf{x}(t_{k-1}) \,|\, \mathbf{y}_{1:k-1})\, \mathrm{d}\mathbf{x}(t_{k-1}). \tag{3.118}$$

3. Update: *Compute posterior distribution of state* $\mathbf{x}(t_k)$ *by the Bayes' rule*

$$p(\mathbf{x}(t_k) \,|\, \mathbf{y}_{1:k}) \propto p(\mathbf{y}_k \,|\, \mathbf{x}(t_k))\, p(\mathbf{x}(t_k) \,|\, \mathbf{y}_{1:k-1}). \tag{3.119}$$

The algorithm above can be used in practice only when the transition density can be solved analytically from the Kolmogorov forward equation. This is the case, for example, in continuous-discrete Kalman filtering model, where the transition density is Gaussian. However, in numerical computations point of view it is more reasonable to avoid computation of the transition density. The filtering algorithm can be equivalently stated as follows:

Algorithm 3.13 (Continuous-discrete Bayesian filtering equations II)**.** *For each measurement* $\mathbf{y}_k$*, do the following:*

1. Prediction: *Solve the predicted probability density* $p(\mathbf{x}(t_k) \mid \mathbf{y}_{1:k-1})$ *from the boundary value problem*

$$\begin{aligned} \frac{\partial p}{\partial t} &= \mathcal{A}_t^*[p] \\ p(\mathbf{x}, t_{k-1}) &= p(\mathbf{x}(t_{k-1}) \mid \mathbf{y}_{1:k-1}) \end{aligned} \tag{3.120}$$

where the predicted density is interpreted as a function of two variables $p(\mathbf{x}(t) \mid \mathbf{y}_{1:k-1}) \triangleq p(\mathbf{x}, t)$ *and the final solution is the function at time* $t = t_k$*.*

2. Update: *Compute posterior distribution of state* $\mathbf{x}(t_k)$ *by the Bayes' rule*

$$p(\mathbf{x}(t_k) \,|\, \mathbf{y}_{1:k}) \propto p(\mathbf{y}_k \,|\, \mathbf{x}(t_k))\, p(\mathbf{x}(t_k) \,|\, \mathbf{y}_{1:k-1}). \tag{3.121}$$

Note that in the first algorithm we actually solve the Green's function of the Kolmogorov forward equation first and then construct the prediction from that. In the second method we directly solve the prediction density from the Kolmogorov forward equation. Neither of these filtering algorithms is often explicitly used in practice, because solving the partial differential equations is impossible in all but simple cases.

Definition 3.3 (Generalized filtering distribution)**.** *The filtering distribution can be defined for all $t \in \mathbb{R}_+$ (not only times t_k) as follows:*

1. *At the measurement times t_k the filtering distribution is the distribution of $\mathbf{x}(t_k)$ given the measurements $\mathbf{y}_{1:k}$:*

$$p(\mathbf{x}(t_k) \,|\, \mathbf{y}_{1:k}) \tag{3.122}$$

2. *At times $t_k < t < t_{k+1}$ the filtering distribution of $\mathbf{x}(t)$ is the distribution obtained from $p(\mathbf{x}(t_k) \,|\, \mathbf{y}_{1:k})$ by prediction to time t:*

$$p(\mathbf{x}(t) \,|\, \mathbf{y}_{1:k}) \int p(\mathbf{x}(t) \,|\, \mathbf{x}(t_k))\, p(\mathbf{x}(t_k) \,|\, \mathbf{y}_{1:k})\, \mathrm{d}\mathbf{x}(t_k), \tag{3.123}$$

This generalized definition of the filtering distribution will be subsequently used in the text.

Example 3.4 (Beneš-Daum filter)**.** *By Equation* (2.82) *in Example 2.2 on page 41, the transition density of the Beneš SDE*

$$\mathrm{d}x = \tanh(x)\,\mathrm{d}t + \mathrm{d}\beta(t), \tag{3.124}$$

is

$$\begin{aligned} p(x(t_k) \,|\, x(t_{k-1})) = {} & \frac{1}{\sqrt{2\pi\Delta t_k}} \frac{\cosh(x(t_k))}{\cosh(x(t_{k-1}))} \exp\left(-\frac{1}{2}\Delta t_k\right) \\ & \times \exp\left(-\frac{1}{2\Delta t_k}(x(t_k) - x(t_{k-1}))^2\right), \end{aligned} \tag{3.125}$$

where $\Delta t_k = t_k - t_{k-1}$. This is also the solution to the associated Kolmogorov forward equation. Assume that the measurements are obtained from the model

$$p(y_k \,|\, x(t_k)) = \mathrm{N}(y_k \,|\, x(t_k), \sigma^2). \tag{3.126}$$

If we assume that the filtering solution at time t_{k-1} is of the form

$$p(x(t_{k-1}) \,|\, y_{1:k-1}) \propto \cosh(x(t_{k-1})) \exp\left(-\frac{1}{2p_{k-1}}(x(t_{k-1}) - m_{k-1})^2\right), \tag{3.127}$$

for some known m_{k-1}, p_{k-1}*, then the Chapman-Kolmogorov equation gives*

$$p(x(t_k) \,|\, y_{1:k-1}) \propto \cosh(x(t_k)) \, \exp\left(-\frac{1}{2p_k^-}(x(t_k) - m_k^-)^2\right). \tag{3.128}$$

where

$$\begin{aligned} m_k^- &= m_{k-1} \\ p_k^- &= p_{k-1} + \Delta t_k. \end{aligned} \tag{3.129}$$

The Bayes' rule gives

$$p(x(t_k) \,|\, y_{1:k}) \propto \cosh(x(t_k)) \, \exp\left(-\frac{1}{2p_k}(x(t_k) - m_k)^2\right). \tag{3.130}$$

where the equations for the parameters (sufficient statistics) can be written as

$$\begin{aligned} m_k &= m_k^- + \left(\frac{p_k^-}{p_k^- + \sigma^2}\right) [y_k - m_k^-] \\ p_k &= p_k^- - \frac{(p_k^-)^2}{p_k^- + \sigma^2}. \end{aligned} \tag{3.131}$$

This result can be easily obtained by completing the squares in the exponent. These prediction and update equations of the Beneš-Daum filter (Daum, 1984) are functionally exactly the same as the equations for a discretely observed Brownian motion, but now the probability density is given by the Equation (3.130). *Particularly, the conditional mean is given as*

$$\mathrm{E}[x(t_k) \mid y_{1:k}] = m_k + P_k \, \tanh(m_k). \tag{3.132}$$

Continuous-Discrete Smoothing Equations

The optimal smoothing of continuous-discrete models is much less discussed in literature than discrete-time smoothing. As in the continuous-discrete filtering case if the transition density of the process can be solved symbolically (see, Algorithm 3.12) then the optimal smoothing can be performed using the discrete-time smoothing equations.

In continuous-time optimal filtering context the optimal smoothing equations are well known and can be expressed in several forms (see, e.g., Liptser and Shiryayev, 1977; Antonelli and Elliott, 1986; Krishnamurthy and Elliott, 2002). The problem in these equations is that they are functions of the measurement process and they are not easy to use in the continuous-discrete filtering case.

Leondes et al. (1970) presents partial differential equations of the smoothing distribution, which are not functions of the measurements directly, but functions of the filtering distributions instead. The disadvantage of these equations is that not only the filtering distribution, but also the first and second order derivatives of the filtering distributions needs to be known. Fortunately, the discrete-time smoothing equations, which also are functions of the filtering distributions only, can be used for computing the smoothing solution of any time instance τ.

Algorithm 3.14 (Continuous-discrete Bayesian smoothing equations)**.** *Given the (generalized) filtering distributions $p(\mathbf{x}(t) \,|\, \mathbf{y}_{1:k})$ the optimal smoothing distribution for any time instance τ such that $t_i < \tau < t_{i+1}$, where t_i and t_{i+1} are the times of the measurements y_i and y_{i+1}, respectively, can be computed as follows:*

1. *Solve the transition densities $p(\mathbf{x}(t_{k+1}) \mid \mathbf{x}(t_k))$ for $k = i+1, \ldots, T$ from the boundary value problem* (3.117)*. Also solve the transition density $p(\mathbf{x}(t_{i+1}) \mid \mathbf{x}(\tau))$.*
2. *Compute the smoothing solution at time step $i+1$, that is, $p(\mathbf{x}(t_{i+1}) \,|\, \mathbf{y}_{1:T})$ using the discrete-time smoothing recursions in the Theorem 3.2*
3. *Compute the smoothed distribution at time instance τ by*

$$\begin{aligned} &p(\mathbf{x}(\tau) \,|\, \mathbf{y}_{1:T}) \\ &\quad = p(\mathbf{x}(\tau) \,|\, \mathbf{y}_{1:k}) \int \left[\frac{p(\mathbf{x}(t_{i+1}) \,|\, \mathbf{x}(\tau))\, p(\mathbf{x}(t_{i+1}) \,|\, \mathbf{y}_{1:T})}{p(\mathbf{x}(t_{i+1}) \,|\, \mathbf{y}_{1:k})} \right] \mathrm{d}\mathbf{x}(t_{i+1}). \end{aligned} \tag{3.133}$$

Note that this equation is simply computation of the discrete-time solution to the additional time instance τ. Although this time instance is not a measurement time instance, the equations are still the same[7]*.*

Continuous-Discrete Kalman Filter

In the *continuous-discrete Kalman filter* (see, e.g., Jazwinski, 1970) the dynamic model is a *linear* stochastic differential equation (see, Section 2.2.6), and the measurements are obtained at discrete instances of time from a linear Gaussian model:

$$\begin{aligned} \mathrm{d}\mathbf{x}(t) &= \mathbf{F}(t)\,\mathbf{x}(t)\,\mathrm{d}t + \mathbf{L}(t)\,\mathrm{d}\boldsymbol{\beta}(t) \\ \mathbf{y}_k &= \mathbf{H}_k\,\mathbf{x}(t_k) + \mathbf{r}_k, \end{aligned} \tag{3.134}$$

where $\mathbf{r}_k \sim \mathrm{N}(\mathbf{0}, \mathbf{R}_k)$, $\mathbf{F}(t)$ and $\mathbf{L}(t)$ are time dependent matrices, and $\boldsymbol{\beta}(t)$ is a Brownian motion with (possibly time varying) diffusion matrix $\mathbf{Q}_c(t)$. The stochastic process $\mathbf{x}(t)$ has the initial distribution $\mathbf{x}(0) \sim \mathrm{N}(\mathbf{m}(0), \mathbf{P}(0))$. As

[7] The proof is not given here, but conceptually we can imagine a measurement with infinite amount of noise, that is, an uninformative measurement to the time instance τ.

shown in Section 2.2.6 the solution $\mathbf{x}(t)$ is a Gaussian process with its mean and covariance given by the differential equations (2.84) and (2.85) in Section 2.2.6. That is, the solution of the corresponding Kolmogorov equation is a Gaussian process, which has the mean and covariance given by those equations.

By using the Theorem 2.8 the continuous-time linear dynamic model in (3.134) can be converted into the *equivalent discrete model*, which is of the same form as the dynamic model of the discrete-time Kalman filter (3.21). This model corresponds to the close form solution to the associated Kolmogorov forward equation.

Algorithm 3.15 (Continuous-discrete Kalman filter I)**.** *The discretization based continuous-discrete Kalman filter can be now expressed as follows:*

- Discretization: *Solve the discrete-time model matrices* $\mathbf{A}_{k-1} \triangleq \mathbf{A}(t_k)$*,* $\mathbf{Q}_{k-1} \triangleq \mathbf{Q}(t_k)$ *from the differential equations*

$$\begin{aligned} \frac{\mathrm{d}\mathbf{A}(t)}{\mathrm{d}t} &= \mathbf{F}(t)\,\mathbf{A}(t) \\ \frac{\mathrm{d}\mathbf{Q}(t)}{\mathrm{d}t} &= \mathbf{F}(t)\,\mathbf{Q}(t) + \mathbf{Q}(t)\,\mathbf{F}^T(t) + \mathbf{L}(t)\,\mathbf{Q}_c(t)\,\mathbf{L}^T(t). \end{aligned} \tag{3.135}$$

  with initial conditions $\mathbf{A}(t_{k-1}) = \mathbf{I}$ *and* $\mathbf{Q}(t_{k-1}) = \mathbf{0}$*. According to the Theorem 2.8 the transition density is now of the form*

$$p(\mathbf{x}(t_k)\,|\,\mathbf{x}(t_{k-1})) = \mathrm{N}(\mathbf{x}(t_k)\,|\,\mathbf{A}_{k-1}\,\mathbf{x}(t_{k-1}), \mathbf{Q}_{k-1}), \tag{3.136}$$

  and thus the discrete-time Kalman filter equations can be applied as such.

- Prediction:

$$\begin{aligned} \mathbf{m}_k^- &= \mathbf{A}_{k-1}\,\mathbf{m}_{k-1} \\ \mathbf{P}_k^- &= \mathbf{A}_{k-1}\,\mathbf{P}_{k-1}\,\mathbf{A}_{k-1}^T + \mathbf{Q}_{k-1}. \end{aligned} \tag{3.137}$$

- Update:

$$\begin{aligned} \mathbf{v}_k &= \mathbf{y}_k - \mathbf{H}_k\,\mathbf{m}_k^- \\ \mathbf{S}_k &= \mathbf{H}_k\,\mathbf{P}_k^-\,\mathbf{H}_k^T + \mathbf{R}_k \\ \mathbf{K}_k &= \mathbf{P}_k^-\,\mathbf{H}_k^T\,\mathbf{S}_k^{-1} \\ \mathbf{m}_k &= \mathbf{m}_k^- + \mathbf{K}_k\,\mathbf{v}_k \\ \mathbf{P}_k &= \mathbf{P}_k^- - \mathbf{K}_k\,\mathbf{S}_k\,\mathbf{K}_k^T. \end{aligned} \tag{3.138}$$

The prediction and update steps above are functionally the same as in discrete-time Kalman filter (repeated here for convenience).

Note that if the dynamic model in (3.134) is *linear and time invariant* (LTI), that is, the matrices $\mathbf{F}$ and $\mathbf{L}$ do not depend on time, the discrete model matrices will depend only on the time difference $\Delta t_{k-1} = t_k - t_{k-1}$, $\mathbf{A}_{k-1} = \mathbf{A}(\Delta t_{k-1})$, $\mathbf{Q}_{k-1} = \mathbf{Q}(\Delta t_{k-1})$. These matrices can be then solved in closed form or by numerical methods (see, Theorem 2.9).

The continuous-discrete Kalman filter corresponding to the second version of the continuous-discrete filter (Algorithm 3.13) is as follows:

Algorithm 3.16 (Continuous-discrete Kalman filter II)**.** *The continuous-discrete Kalman filter consist of the following prediction and update steps:*

- *On* prediction *step, the differential equations*

$$\begin{aligned}\frac{\mathrm{d}\mathbf{m}(t)}{\mathrm{d}t} &= \mathbf{F}(t)\,\mathbf{m}(t)\\ \frac{\mathrm{d}\mathbf{P}(t)}{\mathrm{d}t} &= \mathbf{F}(t)\,\mathbf{P}(t) + \mathbf{P}(t)\,\mathbf{F}^T(t) + \mathbf{L}(t)\,\mathbf{Q}_c(t)\,\mathbf{L}^T(t),\end{aligned} \tag{3.139}$$

  are integrated from the initial conditions $\mathbf{m}(t_{k-1}) = \mathbf{m}_{k-1}$, $\mathbf{P}(t_{k-1}) = \mathbf{P}_{k-1}$ *to time instance* t_k. *The predicted mean and covariance are given as* $\mathbf{m}_k^- = \mathbf{m}(t_k)$ *and* $\mathbf{P}_k^- = \mathbf{P}(t_k)$, *respectively.*

- Update *step is the same as the discrete Kalman filter update step* (3.25) *of Theorem 3.3 (repeated here for convenience):*

$$\begin{aligned}\mathbf{v}_k &= \mathbf{y}_k - \mathbf{H}_k\,\mathbf{m}_k^-\\ \mathbf{S}_k &= \mathbf{H}_k\,\mathbf{P}_k^-\,\mathbf{H}_k^T + \mathbf{R}_k\\ \mathbf{K}_k &= \mathbf{P}_k^-\,\mathbf{H}_k^T\,\mathbf{S}_k^{-1}\\ \mathbf{m}_k &= \mathbf{m}_k^- + \mathbf{K}_k\,\mathbf{v}_k\\ \mathbf{P}_k &= \mathbf{P}_k^- - \mathbf{K}_k\,\mathbf{S}_k\,\mathbf{K}_k^T.\end{aligned} \tag{3.140}$$

In both continuous-discrete Kalman filter formulations the results of filtering are the mean and covariance $\mathbf{m}(t), \mathbf{P}(t)$, which are defined for all t when the filtering result is interpreted in the generalized sense (Definition 3.3). Note that these functions are not continuous at the measurement times. The filtering solution is then of the form

$$p(\mathbf{x}(t)\,|\,\mathbf{y}_{1:k}) = \mathrm{N}(\mathbf{x}(t)\,|\,\mathbf{m}(t), \mathbf{P}(t)), \tag{3.141}$$

where k is such that $t \in [t_k, t_{k+1})$.

Example 3.5 (Kalman filter for Wiener velocity model)**.** *Assume that the state dynamics of a system can be modeled with the Wiener velocity model*

$$\frac{\mathrm{d}^2 x(t)}{\mathrm{d}t^2} = w(t). \tag{3.142}$$

already analyzed in Example 2.3. Further assume that the state $x(t)$ *is observed at discrete instances of time from the model*

$$y_k = x(t_k) + \epsilon_k, \tag{3.143}$$

where $\epsilon_k \sim \mathrm{N}(0, r)$. *The model is now a special case of the continuous-discrete Kalman filtering model* (3.134) *with the following parameter matrices:*

$$\mathbf{F} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \qquad \mathbf{L} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \qquad \mathbf{H}_k = \begin{pmatrix} 1 & 0 \end{pmatrix} \tag{3.144}$$

The spectral density of noise is a scalar $\mathbf{Q}_c = q$ *as is the covariance of the measurement noise* $\mathbf{R}_k = r$. *If we denote the mean and covariance as* $\mathbf{m}(t) = \begin{pmatrix} m_1(t) \\ m_2(t) \end{pmatrix}$ *and* $\mathbf{P}(t) = \begin{pmatrix} P_{11}(t) & P_{12}(t) \\ P_{21}(t) & P_{22}(t) \end{pmatrix}$ *then the component wise prediction differential equations are given as*

$$\begin{aligned}
\mathrm{d}m_1/\mathrm{d}t &= m_2 \\
\mathrm{d}m_2/\mathrm{d}t &= 0 \\
\mathrm{d}P_{11}/\mathrm{d}t &= P_{21} + P_{12} \\
\mathrm{d}P_{12}/\mathrm{d}t &= P_{22} \\
\mathrm{d}P_{21}/\mathrm{d}t &= P_{22} \\
\mathrm{d}P_{22}/\mathrm{d}t &= q.
\end{aligned} \tag{3.145}$$

The same result would be obtained by using the matrices (2.104) *for discrete-time prediction. The update equations are given as*

$$\begin{aligned}
m_{1,k} &= m_{1,k}^- + \left(\frac{P_{11,k}^-}{P_{11,k}^- + r} \right) (y_k - m_{1,k}^-) \\
m_{2,k} &= m_{2,k}^- + \left(\frac{P_{21,k}^-}{P_{11,k}^- + r} \right) (y_k - m_{1,k}^-) \\
P_{11,k} &= P_{11,k}^- - \frac{(P_{11,k}^-)^2}{P_{11,k}^- + r} \\
P_{12,k} &= P_{12,k}^- - \frac{P_{11,k}^- P_{21,k}^-}{P_{11,k}^- + r} \\
P_{21,k} &= P_{21,k}^- - \frac{P_{11,k}^- P_{21,k}^-}{P_{11,k}^- + r} \\
P_{22,k} &= P_{22,k}^- - \frac{(P_{21,k}^-)^2}{P_{11,k}^- + r}.
\end{aligned} \tag{3.146}$$

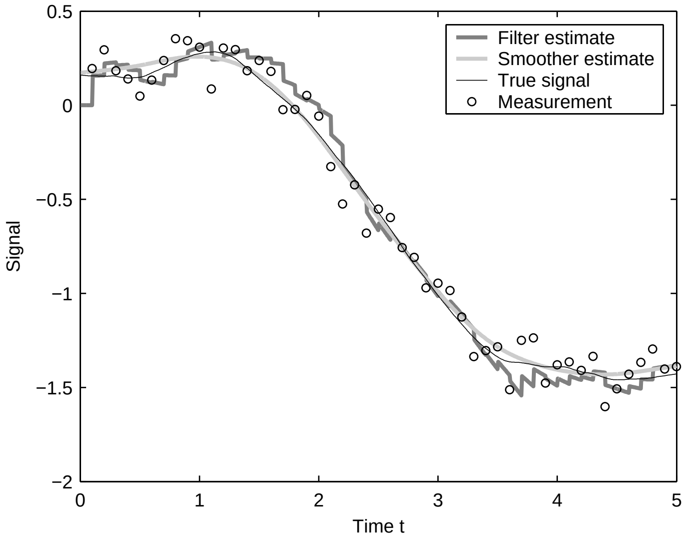

Figure 3.6: Simulated signal, measurements and the estimation results in the continuous-discrete Kalman filtering/smoothing example (Example 3.5/3.6). The signal is a partially observed Wiener velocity model signal.

Continuous-Discrete Kalman Smoother

The continuous-discrete (fixed-interval) Kalman smoother or *Rauch-Tung-Striebel smoother* (Rauch et al., 1965; Gelb, 1974) provides the exact smoothing solution to the Kalman filtering model (3.134). The smoothing solution is a Gaussian process, and the probability density at time t is

$$p(\mathbf{x}(t) \,|\, \mathbf{y}_{1:T}) = \mathrm{N}(\mathbf{x}(t) \,|\, \mathbf{m}^s(t), \mathbf{P}^s(t)), \qquad t_0 \le t \le t_T. \tag{3.147}$$

A straight forward application of the optimal smoothing Algorithm 3.14 means computing the discretization matrices $\mathbf{A}_k$ and $\mathbf{Q}_k$ as in the Algorithm 3.15. Given these matrices the smoothing algorithm reduces to the discrete Kalman smoother in Theorem 3.4. The algorithm can be summarized as follows:

Algorithm 3.17 (Continuous-discrete Kalman smoother I)**.** *The mean $\mathbf{m}^s(\tau)$ and covariance $\mathbf{P}^s(\tau)$ of the smoothed distribution at time instance τ such that $t_i < \tau < t_{i+1}$ can be computed as follows:*

1. *Solve the matrices $\mathbf{A}_k$ and $\mathbf{Q}_k$ for $k = 0, \ldots, T-1$ from the Equation* (3.135)*. Also solve the matrices for $\tau \to t_{i+1}$, here denoted as $\mathbf{A}(\tau)$ and $\mathbf{Q}(\tau)$, in the analogous manner.*

2. *Compute the mean $\mathbf{m}^s(t_{i+1})$ and covariance $\mathbf{P}^s(t_{i+1})$ of the smoothing distribution at time t_{i+1} by using the Kalman smoother Equations* (3.37) *with definitions $\mathbf{m}_k \triangleq \mathbf{m}(t_k)$, $\mathbf{P}_k \triangleq \mathbf{P}(t_k)$, $\mathbf{m}^s_k \triangleq \mathbf{m}^s(t_k)$, $\mathbf{P}^s_k \triangleq \mathbf{P}^s(t_k)$.*

3. *Finally compute*

$$\begin{aligned} \mathbf{C} &= \mathbf{P}(\tau)\,\mathbf{A}^T(\tau)\,[\mathbf{P}(t_{i+1})]^{-1} \\ \mathbf{m}^s(\tau) &= \mathbf{m}(\tau) + \mathbf{C}\,[\mathbf{m}^s(t_{i+1}) - \mathbf{m}(t_{i+1})] \\ \mathbf{P}^s(\tau) &= \mathbf{P}(\tau) + \mathbf{C}\,[\mathbf{P}^s(t_{i+1}) - \mathbf{P}(t_{k+1})]\,\mathbf{C}^T, \end{aligned} \tag{3.148}$$

where $\mathbf{m}(t)$ and $\mathbf{P}(t)$ denote the mean and covariance of the generalized filtering solution at time instance t.

An alternative approach to the smoothing problem is to derive the differential equations for the smoothed solution at every time step as function of the generalized filtering solution. This results in the algorithm given below. These equations are also well known as they already appeared in the original article of Rauch et al. (1965) (see also Gelb, 1974). However, the derivation of these equations is presented in Appendix A.3.2 for completeness and as preparation to the derivation of the continuous-discrete unscented Kalman smoother equations presented later in this section.

Algorithm 3.18 (Continuous-discrete Kalman smoother II)**.** *The differential equations for the smoothed mean $\mathbf{m}^s(t)$ and smoothed covariance $\mathbf{P}^s(t)$ are*

$$\begin{aligned} \frac{\mathrm{d}\mathbf{m}^s(t)}{\mathrm{d}t} &= \mathbf{F}(t)\,\mathbf{m}^s(t) + \mathbf{L}(t)\,\mathbf{Q}_c(t)\,\mathbf{L}^T(t)\,\mathbf{P}^{-1}(t)\;[\mathbf{m}^s(t) - \mathbf{m}(t)] \\ \frac{\mathrm{d}\mathbf{P}^s(t)}{\mathrm{d}t} &= \left[\mathbf{F}(t) + \mathbf{L}(t)\,\mathbf{Q}_c(t)\,\mathbf{L}(t)\,\mathbf{P}^{-1}(t)\right]\;\mathbf{P}^s(t) \\ &\quad + \mathbf{P}^s(t)\;\left[\mathbf{F}(t) + \mathbf{L}(t)\,\mathbf{Q}_c(t)\,\mathbf{L}^T(t)\,\mathbf{P}^{-1}(t)\right]^T - \mathbf{L}(t)\,\mathbf{Q}_c(t)\,\mathbf{L}^T(t), \end{aligned} \tag{3.149}$$

with terminal conditions $\mathbf{m}^s(T) = \mathbf{m}(T), \mathbf{P}^s(T) = \mathbf{P}(T)$. The functions $\mathbf{m}(t)$ and $\mathbf{P}(t)$ are the mean and covariance of the generalized filtering solution.

Example 3.6 (Kalman smoother for Wiener velocity model)**.** *The differential equations of the Kalman smoother for the Wiener velocity model Example 3.5 can be written component wise as*

$$\begin{aligned} \mathrm{d}m^s_1/\mathrm{d}t &= m^s_2 \\ \mathrm{d}m^s_2/\mathrm{d}t &= q\,P^{-1}_{21}\,(m^s_1 - m_1) + q\,P^{-1}_{22}\,(m^s_2 - m_2) \\ \mathrm{d}P^s_{11}/\mathrm{d}t &= P^s_{21} + P^s_{12} \\ \mathrm{d}P^s_{12}/\mathrm{d}t &= P^s_{22} + q\,P^{-1}_{21}\,P^s_{11} + q\,P^s_{12}\,P^{-1}_{22} \\ \mathrm{d}P^s_{21}/\mathrm{d}t &= P^s_{22} + q\,P^{-1}_{21}\,P^s_{11} + q\,P^s_{21}\,P^{-1}_{22} \\ \mathrm{d}P^s_{22}/\mathrm{d}t &= q\,P^{-1}_{21}\,P^s_{12} + 2\,q\,P^{-1}_{22}\,P^s_{22} + q\,P^s_{21}\,P^{-1}_{21} - q, \end{aligned} \tag{3.150}$$

where P_{ij}^{-1} denotes the element ij of the inverse of the covariance matrix $\mathbf{P}(t)$.

3.2.2 Continuous-Discrete Gaussian Approximations

Next Gaussian approximation based methods for non-linear continuous-discrete filtering problems are presented. First the classical Taylor series based continuous-discrete extended Kalman filters and smoothers are presented. Then the novel continuous-discrete unscented Kalman filters and smoothers are presented, which are based on the continuous-time and discrete-time unscented transforms.

The approximate filters presented in this section are based on the second version of the optimal continuous-discrete filter (Algorithm 3.13). In the first order continuous-discrete extended Kalman filter it would be possible to construct a transition density corresponding to the linear approximation of the SDE, but because this does not generalize to the second order continuous-discrete extended Kalman filter or to the continuous-discrete unscented Kalman filter, this discretization based approach is not presented here.

Linear and Quadratic Approximations of SDEs

Next we consider methods for forming Gaussian approximations to the probability density of the state $\mathbf{x}(t)$, which is generated by the stochastic differential equation

$$\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\,\mathrm{d}t + \mathbf{L}(t)\,\mathrm{d}\boldsymbol{\beta}. \tag{3.151}$$

where $\boldsymbol{\beta}(t)$ is a Brownian motion with diffusion matrix $\mathbf{Q}_c(t)$. In Gaussian approximations the idea is to form a Gaussian process, which has the same mean and covariance as the process. The exact differential equations for the mean and covariance are (see, e.g., Jazwinski, 1970; Gelb, 1974)

$$\begin{aligned}
\frac{\mathrm{d}\mathbf{m}(t)}{\mathrm{d}t} &= \mathrm{E}[\mathbf{f}(\mathbf{x}(t), t)] \\
\frac{\mathrm{d}\mathbf{P}(t)}{\mathrm{d}t} &= \mathrm{E}[\mathbf{x}(t)\,\mathbf{f}^T(\mathbf{x}(t), t)] - \mathrm{E}[\mathbf{x}(t)]\ \mathrm{E}[\mathbf{f}^T(\mathbf{x}(t), t)] \\
&\quad + \mathrm{E}[\mathbf{f}(\mathbf{x}(t), t)\,\mathbf{x}^T(t)] - \mathrm{E}[\mathbf{f}(\mathbf{x}(t), t)]\ \mathrm{E}[\mathbf{x}^T(t)] + \mathbf{L}(t)\,\mathbf{Q}_c(t)\,\mathbf{L}(t),
\end{aligned} \tag{3.152}$$

where the expectations are taken with respect to the probability density of $\mathbf{x}(t)$. That is, these expectations could be computed only if we knew the true probability density of $\mathbf{x}(t)$ at all times.

By forming Taylor series expansions of the function $\mathbf{f}(\cdot)$ around the mean it is possible to form approximations to the exact mean and covariance equations (see, e.g., Gelb, 1974, for derivations). These approximations can be formed as follows:

Algorithm 3.19 (Linear approximation of SDE)**.** *When only the first order terms in the Taylor series expansion of* $\mathbf{f}(\cdot)$ *are retained, the following mean and covariance propagation equations are obtained:*

$$\begin{aligned} \frac{\mathrm{d}\mathbf{m}(t)}{\mathrm{d}t} &= \mathbf{f}(\mathbf{m}(t), t) \\ \frac{\mathrm{d}\mathbf{P}(t)}{\mathrm{d}t} &= \mathbf{F}(\mathbf{m}(t), t)\,\mathbf{P}(t) + \mathbf{P}(t)\,\mathbf{F}^T(\mathbf{m}(t), t) + \mathbf{L}(t)\,\mathbf{Q}_c(t)\,\mathbf{L}(t), \end{aligned} \tag{3.153}$$

where the matrix $\mathbf{F}(\cdot)$ *is the Jacobian matrix of the drift term with the elements*

$$F_{ij} = \frac{\partial f_i}{\partial x_j}. \tag{3.154}$$

Algorithm 3.20 (Quadratic approximation of SDE)**.** *Retaining the first and second order terms in the Taylor series expansion of* $\mathbf{f}(\cdot)$ *results in the following equations for the mean and covariance:*

$$\begin{aligned} \frac{\mathrm{d}\mathbf{m}(t)}{\mathrm{d}t} &= \mathbf{f}(\mathbf{m}(t), t) + \frac{1}{2}\partial^2(\mathbf{f}, \mathbf{P}(t)) \\ \frac{\mathrm{d}\mathbf{P}(t)}{\mathrm{d}t} &= \mathbf{F}(\mathbf{m}(t), t)\,\mathbf{P}(t) + \mathbf{P}(t)\,\mathbf{F}^T(\mathbf{m}(t), t) + \mathbf{L}(t)\,\mathbf{Q}_c(t)\,\mathbf{L}(t), \end{aligned} \tag{3.155}$$

where

$$\partial_i^2(\mathbf{f}, \mathbf{P}) = \mathrm{tr}\left\{ \left[\frac{\partial^2 f_i}{\partial x_p \partial x_q} \right] \mathbf{P} \right\}. \tag{3.156}$$

The term in the brackets is the Hessian matrix of f_i*, where the element* pq *is given as* $\partial^2 f_i / (\partial x_p \partial x_q)$ *and* $\mathbf{F}(\cdot)$ *is the Jacobian matrix defined in* (3.154).

Continuous-Discrete Extended Kalman Filter

The continuous-discrete extended Kalman filter (see, e.g., Gelb, 1974) is a classical method for approximating both the discrete and the continuous-discrete nonlinear optimal filters by using Taylor series expansions. The filtering model has the general form

$$\begin{aligned} \mathrm{d}\mathbf{x} &= \mathbf{f}(\mathbf{x}, t)\,\mathrm{d}t + \mathbf{L}(t)\,\mathrm{d}\boldsymbol{\beta}(t) \\ p(\mathbf{y}_k \,|\, \mathbf{x}(t_k)) &= \mathrm{N}(\mathbf{y}_k \,|\, \mathbf{h}(\mathbf{x}(t_k), t_k), \mathbf{R}_k), \end{aligned} \tag{3.157}$$

where

- $\mathbf{x}(t) \in \mathbb{R}^n$ is the state,
- $\mathbf{y}_k \in \mathbb{R}^m$ is the measurement obtained at time instance t_k.
- $\mathbf{f} : \mathbb{R}^n \times \mathbb{R}_+ \mapsto \mathbb{R}^n$ is the drift function,

- $\mathbf{L}(t) \in \mathbb{R}^{n \times s}$ is the (state independent) dispersion matrix,
- $\boldsymbol{\beta}(t) \in \mathbb{R}^{s}$ is Brownian motion with diffusion matrix $\mathbf{Q}_c(t) \in \mathbb{R}^{s \times s}$,
- $\mathbf{h} : \mathbb{R}^n \times \mathbb{R}_+ \mapsto \mathbb{R}^m$ is the measurement model function,
- $\mathbf{R}_k \in \mathbb{R}^{m \times m}$ is the covariance matrix of the measurement $\mathbf{y}_k$.

The first order continuous-discrete extended Kalman filter approximates the dynamic model SDE by the linear approximation given in Algorithm 3.19 and the update step is approximated in the same manner as in discrete-time EKF. Here the term first order means that the dynamic model is approximated with first order (linear) approximation and on update step either the first or second order EKF update step may be used[8].

Algorithm 3.21 (First order continuous-discrete EKF)**.** *The prediction and update steps of the first order continuous-discrete extended Kalman filter are:*

- Prediction. *Integrate the differential equations*

$$\begin{aligned} \frac{\mathrm{d}\mathbf{m}(t)}{\mathrm{d}t} &= \mathbf{f}(\mathbf{m}(t), t) \\ \frac{\mathrm{d}\mathbf{P}(t)}{\mathrm{d}t} &= \mathbf{F}(\mathbf{m}(t), t)\,\mathbf{P}(t) + \mathbf{P}(t)\,\mathbf{F}^T(\mathbf{m}(t), t) + \mathbf{L}(t)\,\mathbf{Q}_c(t)\,\mathbf{L}(t), \end{aligned} \tag{3.158}$$

  where $\mathbf{F}(\cdot)$ is the Jacobian matrix of $\mathbf{f}(\cdot)$ with elements $F_{ij} = \partial f_i / \partial x_j$ from the initial conditions $\mathbf{m}(t_{k-1}) = \mathbf{m}_{k-1}$, $\mathbf{P}(t_{k-1}) = \mathbf{P}_{k-1}$ to the time instance t_k.

- Update. *Perform update step according to the update step of the first order discrete-time EKF in Equation* (3.59) *or the second order discrete-time EKF in Equation* (3.63).

The second order continuous-discrete extended Kalman filter uses the quadratic approximation in Algorithm 3.20 for approximating the non-linear dynamic model.

Algorithm 3.22 (Second order continuous-discrete EKF)**.** *The prediction and update steps of the second order continuous-discrete extended Kalman filter are:*

- Prediction. *Integrate the differential equations*

$$\begin{aligned} \frac{\mathrm{d}\mathbf{m}(t)}{\mathrm{d}t} &= \mathbf{f}(\mathbf{m}(t), t) + \frac{1}{2}\partial^2(\mathbf{f}, \mathbf{P}(t)) \\ \frac{\mathrm{d}\mathbf{P}(t)}{\mathrm{d}t} &= \mathbf{F}(\mathbf{m}(t), t)\,\mathbf{P}(t) + \mathbf{P}(t)\,\mathbf{F}^T(\mathbf{m}(t), t) + \mathbf{L}(t)\,\mathbf{Q}_c(t)\,\mathbf{L}(t), \end{aligned} \tag{3.159}$$

[8] Actually, the UKF update step could be used as well.

where $\mathbf{F}(\cdot)$ *is the Jacobian matrix of* $\mathbf{f}(\cdot)$ *and* $\partial^2(\cdot)$ *is as defined in Equation* (3.156). *The equations are integrated from initial conditions* $\mathbf{m}(t_{k-1}) = \mathbf{m}_{k-1}$, $\mathbf{P}(t_{k-1}) = \mathbf{P}_{k-1}$ *to the time instance* t_k.

- Update. *Perform update step according to the update step of the first order discrete-time EKF in Equation* (3.59) *or the second order discrete-time EKF in Equation* (3.63).

Example 3.7 (CD-EKF Solution to Beneš-Daum filtering problem)**.** *Consider the Beneš-Daum filtering problem presented in Example 3.4*

$$\begin{aligned} \mathrm{d}x &= \tanh(x)\,\mathrm{d}t + \mathrm{d}\beta \\ y_k &= x(t_k) + r_k, \end{aligned} \tag{3.160}$$

where $\beta(t)$ *is a standard Brownian motion (i.e., has the diffusion coefficient* $q = 1$*) and* $r_k \sim \mathrm{N}(0, \sigma^2)$. *The first order continuous-discrete extended Kalman filter equations for this model are*

- Prediction:

$$\begin{aligned} \mathrm{d}m/\mathrm{d}t &= \tanh(m) \\ \mathrm{d}P/\mathrm{d}t &= 2\,(1 - \tanh^2(m))\,P + 1. \end{aligned} \tag{3.161}$$

- Update:

$$\begin{aligned} m_k &= m_k^- + \frac{P_k^-}{(\sigma^2 + P_k^-)}\,(y_k - m_k^-) \\ P_k &= P_k^- - \frac{(P_k^-)^2}{(\sigma^2 + P_k^-)}. \end{aligned} \tag{3.162}$$

Continuous-Discrete Extended Kalman Smoother

The continuous-discrete extended Kalman smoother is an approximate smoothing solution to the filtering model (3.157), which is based on Taylor series approximation of the SDE in the same way as the continuous-discrete extended Kalman filter. The smoother presented here is the one, which results in continuous-time limit of the discrete-time extended Kalman smoother in Algorithm 3.5 and it has the similar form as the continuous-discrete extended Kalman smoother presented (without proof) by Gelb (1974). Higher order forms of the smoother can be found in Leondes et al. (1970) and Sage and Melsa (1971).

Algorithm 3.23 (Continuous-discrete extended Kalman smoother)**.** *The (first order) continuous-discrete extended Kalman smoother uses the linear approximation of SDE in Algorithm 3.19 and the equations are given as*

$$\begin{aligned}\frac{\mathrm{d}\mathbf{m}^s(t)}{\mathrm{d}t} &= \mathbf{f}(\mathbf{m},t) + [\mathbf{F}(\mathbf{m},t)\,\mathbf{P} + \mathbf{L}(t)\,\mathbf{Q}_c(t)\,\mathbf{L}^T(t)]\,\mathbf{P}^{-1}(t)\,[\mathbf{m}^s(t) - \mathbf{m}(t)] \\ \frac{\mathrm{d}\mathbf{P}^s(t)}{\mathrm{d}t} &= [\mathbf{F}(\mathbf{m},t)\,\mathbf{P} + \mathbf{L}(t)\,\mathbf{Q}_c(t)\,\mathbf{L}^T(t)]\,\mathbf{P}^{-1}\,\mathbf{P}^s(t) \\ &\quad + \mathbf{P}^s(t)\,\mathbf{P}^{-1}\,[\mathbf{P}\,\mathbf{F}^T(\mathbf{m},t) + \mathbf{L}(t)\,\mathbf{Q}_c(t)\,\mathbf{L}^T(t)]\; - \mathbf{L}(t)\,\mathbf{Q}_c(t)\,\mathbf{L}^T(t),\end{aligned} \tag{3.163}$$

where $\mathbf{m}(t)$ *and* $\mathbf{P}(t)$ *are the mean and covariance of the (generalized) filtering solution computed by the continuous-discrete extended Kalman filter. The integration is performed backwards starting from the terminal conditions* $\mathbf{m}^s(T) = \mathbf{m}(T)$*,* $\mathbf{P}^s(T) = \mathbf{P}(T)$*.*

Example 3.8 (CD-EKS solution to Beneš-Daum filtering problem)**.** *The equations of the first order extended Kalman smoother for the Beneš-Daum filtering problem in the Example 3.4 are*

$$\begin{aligned}\frac{\mathrm{d}m^s(t)}{\mathrm{d}t} &= \tanh(m(t)) + \left(\frac{(1-\tanh^2(m(t)))\,P(t)+1}{P(t)}\right) \\ &\qquad \times (m^s(t) - m(t)) \\ \frac{\mathrm{d}P^s(t)}{\mathrm{d}t} &= 2\left(\frac{(1-\tanh^2(m(t)))\,P(t)+1}{P(t)}\right)\,P^s(t) - 1,\end{aligned} \tag{3.164}$$

where $m(t)$ *and* $P(t)$ *are the mean and covariance from the continuous-discrete extended Kalman filter.*

Unscented Approximations of SDEs

Next a novel method for forming Gaussian process approximations to stochastic differential equations is presented. The method is based on continuous-time version of the unscented transform.

Theorem 3.5 (Unscented approximation of SDEs)**.** *The continuous-time unscented transform based Gaussian process approximation has the following differential equations for the mean and covariance*

$$\frac{\mathrm{d}\mathbf{m}}{\mathrm{d}t} = \mathbf{f}(\mathbf{X}(t),t)\,\mathbf{w}_m \tag{3.165}$$

$$\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}t} = \mathbf{f}(\mathbf{X}(t),t)\,\mathbf{W}\,\mathbf{X}^T(t) + \mathbf{X}(t)\,\mathbf{W}\,\mathbf{f}^T(\mathbf{X}(t),t) + \mathbf{L}(t)\,\mathbf{Q}_c(t)\,\mathbf{L}^T(t). \tag{3.166}$$

The expression for the sigma point matrix $\mathbf{X}(t)$ is given as[9]

$$\mathbf{X}(t) = \begin{bmatrix}\mathbf{m}(t) & \cdots & \mathbf{m}(t)\end{bmatrix} + \sqrt{c}\,\begin{bmatrix}\mathbf{0} & \sqrt{\mathbf{P}(t)} & -\sqrt{\mathbf{P}(t)}\end{bmatrix}, \tag{3.167}$$

where $\sqrt{\mathbf{P}(t)}$ is a matrix square root of $\mathbf{P}(t)$ (e.g., Cholesky factor) and vector $\mathbf{w}_m$ and matrix $\mathbf{W}$ are defined in (3.81) *and* (3.82)*, respectively.*

Proof. On the small time interval $\Delta t = t_n - t_{n-1}$ the stochastic differential equation (3.151) can be locally approximated to an arbitrary precision with the finite difference[10]

$$\mathbf{x}_n - \mathbf{x}_{n-1} = \mathbf{f}(\mathbf{x}_{n-1}, t_{n-1})\,\Delta t + \mathbf{q}_{n-1} + o(\Delta t), \tag{3.168}$$

where $\mathbf{q}_{n-1}$ is discrete white noise with covariance

$$\mathbf{Q}_{n-1} = \mathbf{L}(t_{n-1})\,\mathbf{Q}_c(t_{n-1})\,\mathbf{L}(t_{n-1})^T\,\Delta t. \tag{3.169}$$

The function $o(\Delta t)$ is such that $o(\Delta t)/\Delta t \to 0$ when $\Delta t \to 0$.

The matrix form unscented transform for the function from $\mathbf{x}_{n-1}$ to $\mathbf{x}_n$ is

$$\begin{aligned}
\mathbf{X}_{n-1} &= \begin{bmatrix}\mathbf{m}_{n-1} & \cdots & \mathbf{m}_{n-1}\end{bmatrix} \\
&\quad + \sqrt{c}\,\begin{bmatrix}\mathbf{0} & \sqrt{\mathbf{P}_{n-1}} & -\sqrt{\mathbf{P}_{n-1}}\end{bmatrix} && (3.170) \\
\mathbf{X}_n &= \mathbf{X}_{n-1} + \mathbf{f}(\mathbf{X}_{n-1}, t_{n-1})\,\Delta t + o(\Delta t) && (3.171) \\
\mathbf{m}_n &= \mathbf{X}_n\,\mathbf{w}_m && (3.172) \\
\mathbf{P}_n &= \mathbf{X}_n\,\mathbf{W}\,[\mathbf{X}_n]^T \\
&\quad + \mathbf{L}(t_{n-1})\,\mathbf{Q}_c(t_{n-1})\,\mathbf{L}^T(t_{n-1})\,\Delta t. && (3.173)
\end{aligned}$$

We can now eliminate the sigma points $\mathbf{X}_n$ from these equations, which gives

$$\begin{aligned}
\mathbf{m}_n &= [\mathbf{X}_{n-1} + \mathbf{f}(\mathbf{X}_{n-1}, t_{n-1})\,\Delta t + o(\Delta t)]\;\mathbf{w}_m \\
&= \mathbf{m}_{n-1} + \mathbf{f}(\mathbf{X}_{n-1}, t_{n-1})\,\Delta t\,\mathbf{w}_m + o(\Delta t),
\end{aligned} \tag{3.174}$$

and

$$\begin{aligned}
\mathbf{P}_n &= [\mathbf{X}_{n-1} + \mathbf{f}(\mathbf{X}_{n-1}, t_{n-1})\,\Delta t + o(\Delta t)]\,\mathbf{W} \\
&\quad \times [\mathbf{X}_{n-1} + \mathbf{f}(\mathbf{X}_{n-1}, t_{n-1})\,\Delta t + o(\Delta t)]^T \\
&\quad + \mathbf{L}(t_{n-1})\,\mathbf{Q}_c(t_{n-1})\,\mathbf{L}^T(t_{n-1})\,\Delta t \\
&= \mathbf{P}_{n-1} + \mathbf{f}(\mathbf{X}_{n-1}, t_{n-1})\,\mathbf{W}\,\mathbf{X}_{n-1}^T\,\Delta t \\
&\quad + \mathbf{X}_{n-1}\,\mathbf{W}\,\mathbf{f}^T(\mathbf{X}_{n-1}, t_{n-1})\,\Delta t \\
&\quad + \mathbf{L}(t_{n-1})\,\mathbf{Q}(t_{n-1})\,\mathbf{L}^T(t_{n-1})\,\Delta t + o(\Delta t),
\end{aligned} \tag{3.175}$$

[9] The sigma points $\mathbf{X}(t)$ act here merely as temporary variables, because they could be completely eliminated by substituting them into the mean and covariance equations.

[10] The analysis is simple, because in this case the Stratonovich and Itô interpretations of the SDE are functionally the same, and the results of normal calculus apply. This could be extended to case $\mathbf{L}(\mathbf{x}, t)$ instead of $\mathbf{L}(t)$, but then more formal analysis based on Itô's formula would be needed.

where we have used the identities $\mathbf{m}_{n-1} = \mathbf{X}_{n-1}\mathbf{w}_m$ and $\mathbf{P}_{n-1} = \mathbf{X}_{n-1}\mathbf{W}\mathbf{X}_{n-1}^T$ and combined all $o(\Delta t)$ terms to one. By rearranging the equations we get

$$\begin{aligned} \frac{\mathbf{m}_n - \mathbf{m}_{n-1}}{\Delta t} &= \mathbf{f}(\mathbf{X}_{n-1}, t_{n-1})\,\mathbf{w}_m \\ &+ o(\Delta t)/\Delta t \qquad (3.176) \\ \frac{\mathbf{P}_n - \mathbf{P}_{n-1}}{\Delta t} &= \mathbf{f}(\mathbf{X}_{n-1}, t_{n-1})\,\mathbf{W}\,\mathbf{X}_{n-1}^T \\ &+ \mathbf{X}_{n-1}\,\mathbf{W}\mathbf{f}^T(\mathbf{X}_{n-1}, t_{n-1}) \\ &+ \mathbf{L}(t_{n-1})\,\mathbf{Q}(t_{n-1})\,\mathbf{L}^T(t_{n-1}) \\ &+ o(\Delta t)/\Delta t, \qquad (3.177) \end{aligned}$$

and in the limit $\Delta t \to 0$ the mean $\mathbf{m}(t)$ and covariance $\mathbf{P}(t)$ satisfy the differential equations in the theorem. □

In the case that the drift term $\mathbf{f}$ in SDE is linear, we have

$$\mathbf{f}(\mathbf{X}(t), t) = \mathbf{F}(t)\,\mathbf{X}(t), \qquad (3.178)$$

and thus

$$\begin{aligned} \mathbf{X}(t)\,\mathbf{W}\,\mathbf{f}^T(\mathbf{X}(t), t) &= \mathbf{X}(t)\,\mathbf{W}\,\mathbf{X}^T(t)\,\mathbf{F}^T(t) \\ &= \mathbf{P}(t)\,\mathbf{F}^T(t) \qquad (3.179) \\ \mathbf{f}(\mathbf{X}(t), t)\,\mathbf{w}_m &= \mathbf{F}(t)\,\mathbf{X}(t)\,\mathbf{w}_m \\ &= \mathbf{F}(t)\,\mathbf{m}(t). \qquad (3.180) \end{aligned}$$

The differential equations now reduce to

$$\frac{\mathrm{d}\mathbf{m}(t)}{\mathrm{d}t} = \mathbf{F}(t)\,\mathbf{m}(t) \qquad (3.181)$$

$$\frac{\mathrm{d}\mathbf{P}(t)}{\mathrm{d}t} = \mathbf{F}(t)\,\mathbf{P}(t) + \mathbf{P}(t)\,\mathbf{F}^T(t) + \mathbf{L}(t)\,\mathbf{Q}_c(t)\,\mathbf{L}^T(t), \qquad (3.182)$$

which are the prediction equations of the continuous-discrete Kalman filter.

Continuous-Discrete Unscented Kalman Filter

The dynamic model of the continuous-discrete filtering model can also be approximated using the unscented approximation of SDE in Algorithm 3.5. This results in the novel *continuous-discrete unscented Kalman filter*, which will be presented next. The filter can be applied to models of the similar form as the continuous-discrete EKF, that is, to models of the form (3.157).

Algorithm 3.24 (Continuous-discrete unscented Kalman filter)**.** *The continuous-discrete unscented Kalman filter consist of the following prediction and update steps:*

- Prediction. *Integrate the differential equations*

$$\begin{aligned}\mathbf{X}(t) &= \begin{bmatrix}\mathbf{m}(t) & \cdots & \mathbf{m}(t)\end{bmatrix} + \sqrt{c}\,\begin{bmatrix}\mathbf{0} & \sqrt{\mathbf{P}(t)} & -\sqrt{\mathbf{P}(t)}\end{bmatrix} \\ \frac{\mathrm{d}\mathbf{m}}{\mathrm{d}t} &= \mathbf{f}(\mathbf{X}(t),t)\,\mathbf{w}_m \\ \frac{\mathrm{d}\mathbf{P}}{\mathrm{d}t} &= \mathbf{f}(\mathbf{X}(t),t)\,\mathbf{W}\,\mathbf{X}^T(t) + \mathbf{X}(t)\,\mathbf{W}\,\mathbf{f}^T(\mathbf{X}(t),t) + \mathbf{L}(t)\,\mathbf{Q}_c(t)\,\mathbf{L}^T(t).\end{aligned} \tag{3.183}$$

from the initial conditions $\mathbf{m}(t_{k-1}) = \mathbf{m}_{k-1}$, $\mathbf{P}(t_{k-1}) = \mathbf{P}_{k-1}$ *to time instance* t_k. *The predicted mean and covariance are given as* $\mathbf{m}_k^- = \mathbf{m}(t_k)$ *and* $\mathbf{P}_k^- = \mathbf{P}(t_k)$*, respectively.*

- Update. *The update step is the same as the discrete-time unscented Kalman filter update step* (3.90) *of Algorithm 3.7 (repeated here for convenience):*

$$\begin{aligned}\mathbf{X}_k^- &= \begin{bmatrix}\mathbf{m}_k^- & \cdots & \mathbf{m}_k^-\end{bmatrix} + \sqrt{c}\begin{bmatrix}\mathbf{0} & \sqrt{\mathbf{P}_k^-} & -\sqrt{\mathbf{P}_k^-}\end{bmatrix} \\ \mathbf{Y}_k^- &= \mathbf{h}(\mathbf{X}_k^-, k) \\ \boldsymbol{\mu}_k &= \mathbf{Y}_k^-\,\mathbf{w}_m \\ \mathbf{S}_k &= \mathbf{Y}_k^-\,\mathbf{W}\,[\mathbf{Y}_k^-]^T + \mathbf{R}_k \\ \mathbf{C}_k &= \mathbf{X}_k^-\,\mathbf{W}\,[\mathbf{Y}_k^-]^T \\ \mathbf{K}_k &= \mathbf{C}_k\,\mathbf{S}_k^{-1} \\ \mathbf{m}_k &= \mathbf{m}_k^- + \mathbf{K}_k\,[\mathbf{y}_k - \boldsymbol{\mu}_k] \\ \mathbf{P}_k &= \mathbf{P}_k^- - \mathbf{K}\,\mathbf{S}_k\,\mathbf{K}_k^T.\end{aligned} \tag{3.184}$$

Example 3.9 (CD-UKF solution to Beneš-Daum filtering problem)**.** *The continuous-discrete unscented Kalman filter equations for the Beneš-Daum filtering problem in Example 3.4 are*

- Prediction.

$$\begin{aligned}\mathbf{X} &= \begin{pmatrix}m & m+\sqrt{c\,P} & m-\sqrt{c\,P}\end{pmatrix} \\ \mathrm{d}m/\mathrm{d}t &= \tanh(\mathbf{X})\,\mathbf{w}_m \\ \mathrm{d}P/\mathrm{d}t &= \tanh(\mathbf{X})\,\mathbf{W}\,\mathbf{X}^T + \mathbf{X}\,\mathbf{W}\,\tanh(\mathbf{X})^T + 1,\end{aligned} \tag{3.185}$$

where $c = 0.75$ *and*

$$\mathbf{w}_m = \begin{pmatrix}-0.3333 \\ 0.6667 \\ 0.6667\end{pmatrix} \qquad \mathbf{W} = \begin{pmatrix}4.4444 & -2.2222 & -2.2222 \\ -2.2222 & 1.4444 & 0.7778 \\ -2.2222 & 0.7778 & 1.4444\end{pmatrix}. \tag{3.186}$$

- Update. *Because the measurement model is linear the update step is the same as in EKF solution and it is not repeated here.*

Continuous-Discrete Unscented Kalman Smoother

Using the same idea as in the derivation of the continuous-discrete Kalman smoother in Appendix A.3.2 it is also possible to find the continuous-time limit of the unscented Kalman smoother equations. This derivation is presented in Appendix A.3.4 and the result is the following novel algorithm:

Algorithm 3.25 (Continuous-discrete unscented Kalman smoother)**.** *The continuous-discrete unscented Kalman smoother equations are given as*

$$\begin{aligned}\frac{\mathrm{d}\mathbf{m}^s(t)}{\mathrm{d}t} &= \mathbf{f}(\mathbf{X},t)\,\mathbf{w}_m \\ &\quad + [\mathbf{f}(\mathbf{X},t)\,\mathbf{W}\,\mathbf{X}^T(t) + \mathbf{L}(t)\,\mathbf{Q}_c(t)\,\mathbf{L}^T(t)]\,\mathbf{P}^{-1}(t)\,[\mathbf{m}^s(t) - \mathbf{m}(t)] \\ \frac{\mathrm{d}\mathbf{P}^s(t)}{\mathrm{d}t} &= [\mathbf{f}(\mathbf{X},t)\,\mathbf{W}\,\mathbf{X}^T(t) + \mathbf{L}(t)\,\mathbf{Q}_c(t)\,\mathbf{L}^T(t)]\,\mathbf{P}^{-1}\,\mathbf{P}^s(t) \\ &\quad + \mathbf{P}^s(t)\,\mathbf{P}^{-1}\,[\mathbf{X}(t)\,\mathbf{W}\,\mathbf{f}^T(\mathbf{X},t) + \mathbf{L}(t)\,\mathbf{Q}_c(t)\,\mathbf{L}^T(t)] \\ &\quad - \mathbf{L}(t)\,\mathbf{Q}_c(t)\,\mathbf{L}^T(t),\end{aligned} \tag{3.187}$$

where the sigma points $\mathbf{X}(t)$ are the sigma points of the continuous-discrete unscented Kalman filter, $\mathbf{P}(t) = \mathbf{X}(t)\,\mathbf{W}\,\mathbf{X}^T(t)$ and $\mathbf{m}(t) = \mathbf{X}(t)\,\mathbf{w}_m$. The integration is performed backwards starting from the terminal conditions $\mathbf{m}^s(T) = \mathbf{m}(T)$, $\mathbf{P}^s(T) = \mathbf{P}(T)$.

If the drift term is linear $\mathbf{f}(\mathbf{x},t) = \mathbf{F}\,\mathbf{x}$, these smoother equations can be seen to reduce to the Rauch-Tung-Striebel smoother equations.

Example 3.10 (CD-UKS solution to Beneš-Daum filtering problem)**.** *The continuous-discrete unscented Kalman smoother solution to the Beneš-Daum filtering problem in Example 3.4 can be written as*

$$\begin{aligned}\mathbf{F}(t) &= \tanh(\mathbf{X}(t)) \\ \frac{\mathrm{d}m^s(t)}{\mathrm{d}t} &= \mathbf{F}(t)\,\mathbf{w}_m + \left(\frac{\mathbf{F}(t)\,\mathbf{W}\,\mathbf{X}^T(t) + 1}{\mathbf{X}(t)\,\mathbf{W}\,\mathbf{X}^T(t)}\right) \\ &\qquad \times (m^s(t) - \mathbf{X}(t)\,\mathbf{w}_m) \\ \frac{\mathrm{d}P^s(t)}{\mathrm{d}t} &= 2\left(\frac{\mathbf{F}(t)\,\mathbf{W}\,\mathbf{X}^T(t) + 1}{\mathbf{X}(t)\,\mathbf{W}\,\mathbf{X}^T(t)}\right)P^s(t) - 1.\end{aligned} \tag{3.188}$$

3.2.3 Continuous-Discrete Sequential Importance Resampling

This section presents sequential importance resampling (i.e., particle filtering) based methods for optimal estimation of continuous-discrete filtering models. First the bootstrap filter solution is provided, which is conceptually simple method for approximating the optimal solution. Then novel measure transform based methods for more general continuous-discrete sequential importance resampling are presented.

Continuous-Discrete Bootstrap Filter

A bootstrap filter can be very easily implemented for the continuous-discrete filtering problems of the general form

$$\begin{aligned} \mathrm{d}\mathbf{x} &= \mathbf{f}(\mathbf{x},t)\,\mathrm{d}t + \mathbf{L}(\mathbf{x},t)\,\mathrm{d}\boldsymbol{\beta} \\ \mathbf{y}_k &\sim p(\mathbf{y}_k \,|\, \mathbf{x}(t_k)), \end{aligned} \tag{3.189}$$

because samples from the transition density of the dynamic model can be easily generated by numerically simulating the stochastic differential equation (see, Section 2.2.7).

Algorithm 3.26 (Continuous-discrete bootstrap filter)**.** *Bootstrap filtering for discretely observed stochastic differential equation can be performed as follows:*

1. *Simulate trajectories* $\{\mathbf{x}^{(i)}(t) \;:\; t_{k-1} \le t \le t_k, i = 1,...,N\}$ *from the equation*

$$\begin{aligned} \mathrm{d}\mathbf{x}^{(i)} &= \mathbf{f}(\mathbf{x}^{(i)},t)\,\mathrm{d}t + \mathbf{L}(\mathbf{x}^{(i)},t)\,\mathrm{d}\boldsymbol{\beta}^{(i)}(t) \\ \mathbf{x}^{(i)}(t_{k-1}) &= \mathbf{x}^{(i)}_{k-1}. \end{aligned} \tag{3.190}$$

   with independent Brownian motions $\boldsymbol{\beta}^{(i)}(t)$ *and set* $\mathbf{x}^{(i)}_k = \mathbf{x}^{(i)}(t_k)$*. Now each* $\mathbf{x}^{(i)}_k$ *is a random draw from the transition distribution* $p(\mathbf{x}_k \,|\, \mathbf{x}^{(i)}_{k-1})$.

2. *Compute the new weights*

$$w^{(i)}_k \propto w^{(i)}_{k-1}\, p(\mathbf{y}_k \,|\, \mathbf{x}^{(i)}_k). \tag{3.191}$$

3. *Resample by choosing from* $\{\mathbf{x}^{(i)}_k \;:\; i = 1,...,N\}$ *with the probabilities given by the weights.*

The bootstrap filter suffers from the problem that using the dynamic model as the importance distribution is not very efficient, and the bootstrap filter is likely to produce degenerate approximations if the dynamic model is not very accurate. In the next sections it is shown how more efficient importance distributions (or actually importance processes) can be used in the continuous-discrete filtering problem.

Sequential Importance Resampling Filter for Absolutely Continuous SDEs

Now the sequential importance resampling of a restricted class SDEs is considered. This will serve as the basis for the sequential importance resampling of more general SDEs. First the sequential importance resampling algorithm for the restricted class of SDEs in the following form is derived:

$$\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\,\mathrm{d}t + \mathbf{L}\,\mathrm{d}\boldsymbol{\beta}, \tag{3.192}$$

where $\mathbf{L}$ is time independent and invertible.

Assume that there exists importance process with the equation

$$\mathrm{d}\mathbf{s} = \mathbf{g}(\mathbf{s}, t)\,\mathrm{d}t + \mathbf{B}\,\mathrm{d}\boldsymbol{\beta}, \tag{3.193}$$

where $\mathbf{B}$ is time independent and invertible. Because the matrices $\mathbf{L}$ and $\mathbf{B}$ are invertible, the probability measures of $\mathbf{x}$ and $\mathbf{s}$ are absolutely continuous with respect to the probability measure of the driving Brownian motion $\boldsymbol{\beta}$. The likelihood ratio of the processes can be computed as follows:

Theorem 3.6 (Likelihood ratio of SDEs I)**.** *Assume that the processes $\mathbf{x}(t)$ and $\mathbf{s}(t)$ are generated by the stochastic differential equations*

$$\begin{aligned}
\mathrm{d}\mathbf{x} &= \mathbf{f}(\mathbf{x}, t)\,\mathrm{d}t + \mathbf{L}\,\mathrm{d}\boldsymbol{\beta}, & \mathbf{x}(0) &= \mathbf{x}_0 & (3.194)\\
\mathrm{d}\mathbf{s} &= \mathbf{g}(\mathbf{s}, t)\,\mathrm{d}t + \mathbf{B}\,\mathrm{d}\boldsymbol{\beta}, & \mathbf{s}(0) &= \mathbf{x}_0, & (3.195)
\end{aligned}$$

where $\mathbf{f}(\mathbf{x}, t)$ and $\mathbf{g}(\mathbf{s}, t)$ are bounded and measurable, $\mathbf{L}$ and $\mathbf{B}$ are invertible matrices and $\boldsymbol{\beta}(t)$ is a Brownian motion with respect to measure P. Then the expectations of $\mathbf{x}(t)$ under measure P can be expressed as

$$\mathrm{E}[\mathbf{h}(\mathbf{x}(t))] = \mathrm{E}[Z(t;\omega)\,\mathbf{h}(\mathbf{s}^*(t))], \tag{3.196}$$

where the scaled version of the process $\mathbf{s}(t)$ is defined as

$$\mathbf{s}^*(t) = \mathbf{x}_0 + \mathbf{L}\,\mathbf{B}^{-1}(\mathbf{s}(t) - \mathbf{x}_0), \tag{3.197}$$

and the likelihood ratio $(\mathrm{d}P_x/\mathrm{d}P_{s*})(t;\omega) = Z(t;\omega)$ is

$$\begin{aligned}
Z(t;\omega) = \exp\Bigg(&\int_0^t \left[\mathbf{L}^{-1}\,\mathbf{f}(\mathbf{s}^*(t), t) - \mathbf{B}^{-1}\,\mathbf{g}(\mathbf{s}(t), t)\right]^T\,\mathrm{d}\boldsymbol{\beta}(t;\omega)\\
&+ \int_0^t [\mathbf{L}^{-1}\,\mathbf{f}(\mathbf{s}^*(t), t)]^T\,[\mathbf{B}^{-1}\,\mathbf{g}(\mathbf{s}(t), t)]\,\mathrm{d}t\\
&- \frac{1}{2}\int_0^t \left(||\mathbf{L}^{-1}\,\mathbf{f}(\mathbf{s}^*(t), t)||^2 + ||\mathbf{B}^{-1}\,\mathbf{g}(\mathbf{s}(t), t)||^2\right)\mathrm{d}t\Bigg).
\end{aligned} \tag{3.198}$$

Proof. Define

$$\begin{aligned}Z'(t;\omega) = \exp\Bigg(&- \int_0^t \left[\mathbf{B}^{-1}\,\mathbf{g}(\mathbf{s}(t),t)\right]^T \,\mathrm{d}\boldsymbol{\beta}(t;\omega) \\ &- \frac{1}{2}\int_0^t ||\mathbf{B}^{-1}\,\mathbf{g}(\mathbf{s}(t),t)||^2 \,\mathrm{d}t\Bigg),\end{aligned} \tag{3.199}$$

then by Theorem 2.5 $\mathbf{n}(t) = \mathbf{B}^{-1}\,(\mathbf{s}(t) - \mathbf{x}_0)$ is a standard Brownian motion with respect to measure $P'(\mathrm{d}\omega) = Z'(t;\omega)\,P(\mathrm{d}\omega)$. Now define

$$\begin{aligned}Z''(t;\omega) = \exp\Bigg(&\int_0^t \left[\mathbf{L}^{-1}\,\mathbf{f}(\mathbf{x}_0 + \mathbf{L}\,\mathbf{n}(t;\omega),t)\right]^T \,\mathrm{d}\mathbf{n}(t;\omega) \\ &- \frac{1}{2}\int_0^t ||\mathbf{L}^{-1}\,\mathbf{f}(\mathbf{x}_0 + \mathbf{L}\,\mathbf{n}(t;\omega),t)||^2 \,\mathrm{d}t\Bigg)\end{aligned} \tag{3.200}$$

By Theorem 2.4 the expectation of function $\mathbf{h}(\mathbf{x}(t))$ with respect to the measure $P'(\mathrm{d}\omega)$ can be expressed as

$$\begin{aligned}\mathrm{E}'[\mathbf{h}(\mathbf{x}(t))] &= \mathrm{E}'[Z''(t;\omega)\,\mathbf{h}(\mathbf{x}_0 + \mathbf{L}\,\mathbf{n}(t))] \\ &= \mathrm{E}'[Z''(t;\omega)\,\mathbf{h}(\mathbf{x}_0 + \mathbf{L}\,\mathbf{B}^{-1}\,(\mathbf{s}(t) - \mathbf{x}_0)] \\ &= \mathrm{E}'[Z''(t;\omega)\,\mathbf{h}(\mathbf{s}^*(t))].\end{aligned} \tag{3.201}$$

Because the likelihood ratio between the measures $P'(\mathrm{d}\omega)$ and $P(\mathrm{d}\omega)$ is $Z'(t;\omega)$, the expectation can be written as

$$\mathrm{E}[\mathbf{h}(\mathbf{x}(t))] = \mathrm{E}[Z(t;\omega)\,\mathbf{h}(\mathbf{s}^*(t))], \tag{3.202}$$

where

$$\begin{aligned}
Z(t;\omega) &= Z''(t;\omega)Z'(t;\omega) \\
&= \exp\Bigg(-\int_0^t \left[\mathbf{B}^{-1}\,\mathbf{g}(\mathbf{s}(t),t)\right]^T \,\mathrm{d}\boldsymbol{\beta}(t;\omega) \\
&\qquad -\frac{1}{2}\int_0^t ||\mathbf{B}^{-1}\,\mathbf{g}(\mathbf{s}(t),t)||^2\,\mathrm{d}t\Bigg) \\
&\times \exp\Bigg(\int_0^t \left[\mathbf{L}^{-1}\,\mathbf{f}(\mathbf{x}_0+\mathbf{L}\,\mathbf{n}(t;\omega),t)\right]^T \,\mathrm{d}\mathbf{n}(t;\omega) \\
&\qquad -\frac{1}{2}\int_0^t ||\mathbf{L}^{-1}\,\mathbf{f}(\mathbf{x}_0+\mathbf{L}\,\mathbf{n}(t;\omega),t)||^2\,\mathrm{d}t,\Bigg) \\
&= \exp\Bigg(-\int_0^t \left[\mathbf{B}^{-1}\,\mathbf{g}(\mathbf{s}(t),t)\right]^T \,\mathrm{d}\boldsymbol{\beta}(t;\omega) \\
&\qquad -\frac{1}{2}\int_0^t ||\mathbf{B}^{-1}\,\mathbf{g}(\mathbf{s}(t),t)||^2\,\mathrm{d}t\Bigg) \\
&\times \exp\Bigg(\int_0^t \left[\mathbf{L}^{-1}\,\mathbf{f}(\mathbf{s}^*(t),t)\right]^T \,\mathbf{B}^{-1}\left[\mathbf{g}(\mathbf{s},t)\,\mathrm{d}t+\mathbf{B}\,\mathrm{d}\boldsymbol{\beta}(t)\right] \\
&\qquad -\frac{1}{2}\int_0^t ||\mathbf{L}^{-1}\,\mathbf{f}(\mathbf{s}^*(t),t)||^2\,\mathrm{d}t,\Bigg) \\
&= \exp\Bigg(\int_0^t \left[\mathbf{L}^{-1}\,\mathbf{f}(\mathbf{s}^*(t),t)-\mathbf{B}^{-1}\,\mathbf{g}(\mathbf{s}(t),t)\right]^T \,\mathrm{d}\boldsymbol{\beta}(t;\omega) \\
&\qquad +\int_0^t [\mathbf{L}^{-1}\,\mathbf{f}(\mathbf{s}^*(t),t)]^T\,[\mathbf{B}^{-1}\,\mathbf{g}(\mathbf{s},t)]\,\mathrm{d}t \\
&\qquad -\frac{1}{2}\int_0^t \left(||\mathbf{L}^{-1}\,\mathbf{f}(\mathbf{s}^*(t),t)||^2+||\mathbf{B}^{-1}\,\mathbf{g}(\mathbf{s}(t),t)||^2\right)\mathrm{d}t\Bigg).
\end{aligned} \tag{3.203}$$

□

Theorem 3.6 actually states that given a set of samples from the process $\mathbf{s}(t)$ we can form a set of importance samples from $\mathbf{x}(t)$ by scaling $\mathbf{s}(t)$, computing the corresponding values $Z(t;\omega)$ and using them as the importance weights. The values can be computed by using any numerical integration method as long as the method approximates the strong solution, not only a weak solution. The strong solution is needed to ensure that the weights $Z(t;\omega)$ are adapted to the same Brownian motion as $\mathbf{s}(t)$.

Algorithm 3.27 (Importance sampling of SDE I)**.** *Using the process $\mathbf{s}(t)$ as the importance process, weighted sample from $\mathbf{x}(t)$ at time T can be generated as follows:*

1. *Randomly draw N Brownian motions $\{\boldsymbol{\beta}^{(i)}(t), 0 \le t \le T, i = 1, \ldots, N\}$ and simulate the corresponding (scaled) importance processes*

$$\mathrm{d}\mathbf{s}^{(i)} = \mathbf{g}(\mathbf{s}^{(i)}, t)\,\mathrm{d}t + \mathbf{B}\,\mathrm{d}\boldsymbol{\beta}^{(i)}, \qquad \mathbf{s}^{(i)}(0) = \mathbf{x}_0 \tag{3.204}$$

from $t = 0$ to $t = T$, and compute

$$\mathbf{s}^{*(i)}(t) = \mathbf{x}_0 + \mathbf{L}\,\mathbf{B}^{-1}\,(\mathbf{s}(t) - \mathbf{x}_0), \tag{3.205}$$

and set

$$\mathbf{x}^{(i)} = \mathbf{s}^{*(i)}(T). \tag{3.206}$$

2. *For each i compute*

$$\begin{aligned} w^{(i)} = \exp\Bigg(&\int_0^T \Big[\mathbf{L}^{-1}\,\mathbf{f}(\mathbf{s}^{*(i)}(t), t) - \mathbf{B}^{-1}\,\mathbf{g}(\mathbf{s}^{(i)}(t), t)\Big]^T\,\mathrm{d}\boldsymbol{\beta}^{(i)} \\ &+ \int_0^T [\mathbf{L}^{-1}\,\mathbf{f}(\mathbf{s}^{*(i)}(t), t)]^T\,[\mathbf{B}^{-1}\,\mathbf{g}(\mathbf{s}^{(i)}(t), t)]\,\mathrm{d}t \\ &- \frac{1}{2}\int_0^T \Big(||\mathbf{L}^{-1}\,\mathbf{f}(\mathbf{s}^{*(i)}(t), t)||^2 + ||\mathbf{B}^{-1}\,\mathbf{g}(\mathbf{s}^{(i)}(t), t)||^2\Big)\,\mathrm{d}t\Bigg). \end{aligned} \tag{3.207}$$

3. *Now $\{(\mathbf{x}^{(i)}, w^{(i)}) \,:\, i = 1, \ldots, N\}$ is a set of importance samples such that for any function $\mathbf{h}(\cdot)$*

$$\mathrm{E}[\mathbf{h}(\mathbf{x}(T))] \approx \sum_i w^{(i)}\,\mathbf{h}(\mathbf{x}^{(i)}), \tag{3.208}$$

where $\mathbf{x}(T)$ is the solution to the stochastic differential equation

$$\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\,\mathrm{d}t + \mathbf{L}\,\mathrm{d}\boldsymbol{\beta}, \qquad \mathbf{x}(0) = \mathbf{x}_0 \tag{3.209}$$

at time T.

Assume that the filtering model is of the form

$$\begin{aligned} \mathrm{d}\mathbf{x} &= \mathbf{f}(\mathbf{x}, t)\,\mathrm{d}t + \mathbf{L}\,\mathrm{d}\boldsymbol{\beta} \\ \mathbf{y}_k &\sim p(\mathbf{y}_k \,|\, \mathbf{x}(t_k)), \end{aligned} \tag{3.210}$$

where $\mathbf{L}$ is an invertible matrix. Further assume that there exists importance process $\mathbf{s}(t)$, which is defined by the SDE

$$\mathrm{d}\mathbf{s} = \mathbf{g}(\mathbf{s}, t)\,\mathrm{d}t + \mathbf{B}\,\mathrm{d}\boldsymbol{\beta}, \tag{3.211}$$

and which has the law that is a rough approximation to the filtering (or smoothing) result of the model (3.210), at least at the measurement times. The matrix $\mathbf{B}$ is also assumed to be invertible.

Now it is possible to generate a set of importance samples from the conditioned (i.e. filtered) process $\mathbf{x}(t)$, which is conditional to the measurements $\mathbf{y}_{1:k}$ using $\mathbf{s}(t)$ as the importance process. The motivation of this is that because the process $\mathbf{s}(t)$ is already an approximation to the optimal result, using it as the importance process is likely to reduce the degeneracy problem in the bootstrap filter.

Because the measures of both the processes are absolutely continuous with respect to the measures of the driving Brownian motions it is possible to use the Algorithm 3.27 for generating the importance samples. The continuous-discrete SIR filter for the model can be now constructed with a slight modification to the discrete-time SIR (Algorithm 3.9) as follows:

Algorithm 3.28 (Continuous-discrete SIR I)**.** *Given the importance process* $\mathbf{s}(t)$*, a weighted set of samples* $\{\mathbf{x}_{k-1}^{(i)}, w_{k-1}^{(i)}\}$ *and the new measurement* $\mathbf{y}_k$*, a single step of continuous-discrete sequential importance resampling can be now performed as follows:*

1. *Draw* N *Brownian motions* $\{\boldsymbol{\beta}^{(i)}(t), t_{k-1} \le t \le t_k, i = 1, \ldots, N\}$ *and simulate the corresponding importance processes*

$$\mathrm{d}\mathbf{s}^{(i)} = \mathbf{g}(\mathbf{s}^{(i)}, t)\,\mathrm{d}t + \mathbf{B}\,\mathrm{d}\boldsymbol{\beta}^{(i)}, \qquad \mathbf{s}^{(i)}(t_{k-1}) = \mathbf{x}_{k-1}^{(i)} \tag{3.212}$$

from $t = t_{k-1}$ *to* $t = t_k$*, and compute*

$$\mathbf{s}^{*(i)}(t) = \mathbf{x}_{k-1}^{(i)} + \mathbf{L}\,\mathbf{B}^{-1}\,(\mathbf{s}^{(i)}(t) - \mathbf{x}_{k-1}^{(i)}), \tag{3.213}$$

and set

$$\mathbf{x}_k^{(i)} = \mathbf{s}^{*(i)}(t_k). \tag{3.214}$$

2. *For each i compute*

$$\begin{aligned} w_k^{(i)} = w_{k-1}^{(i)} \exp\Bigg(&\int_{t_{k-1}}^{t_k} \left[\mathbf{L}^{-1}\,\mathbf{f}(\mathbf{s}^{*(i)}(t),t) - \mathbf{B}^{-1}\,\mathbf{g}(\mathbf{s}^{(i)}(t),t)\right]^T \mathrm{d}\boldsymbol{\beta}^{(i)} \\ &+ \int_{t_{k-1}}^{t_k} [\mathbf{L}^{-1}\,\mathbf{f}(\mathbf{s}^{*(i)}(t),t)]^T\,[\mathbf{B}^{-1}\,\mathbf{g}(\mathbf{s}^{(i)}(t),t)]\,\mathrm{d}t \\ &- \frac{1}{2}\int_{t_{k-1}}^{t_k} \left(||\mathbf{L}^{-1}\,\mathbf{f}(\mathbf{s}^{*(i)}(t),t)||^2 + ||\mathbf{B}^{-1}\,\mathbf{g}(\mathbf{s}^{(i)}(t),t)||^2\right)\mathrm{d}t\Bigg) \\ &\times p(\mathbf{y}_k \,|\, \mathbf{x}_k^{(i)}). \end{aligned} \tag{3.215}$$

and re-normalize the weights to sum to unity.

3. *If the effective number of weights* (3.97) *is too low, perform resampling.*

Example 3.11 (CD-SIR solution to Beneš-Daum filtering problem)**.** *Consider the Beneš-Daum filtering problem presented in Example 3.4, which is of the form*

$$\begin{aligned} \mathrm{d}x &= \tanh(x)\,\mathrm{d}t + \mathrm{d}\beta \\ y_k &= x(t_k) + r_k, \qquad r_k \sim \mathrm{N}(0,\sigma^2). \end{aligned} \tag{3.216}$$

Assume that sampling interval is 1. If the state at time t_{k-1} is x_{k-1} and the y_k is measurement at time t_k, then a reasonable importance process can be formed by using either the EKF in Example 3.7 or the UKF in Example 3.9. This results in Gaussian approximation with mean m_k and variance P_k.

A process, which starts at x_{k-1} and has the above mean and covariance at $t_k = t_{k-1} + 1$ is

$$\mathrm{d}s = (m_k - x_{k-1})\,\mathrm{d}t + \sqrt{P_k}\,\mathrm{d}\beta \qquad s(t_{k-1}) = x_{k-1}, \tag{3.217}$$

and this process can be used as the importance process. The scaled importance process now has the equation

$$\mathrm{d}s^* = \left(\frac{m_k - x_{k-1}}{\sqrt{P_k}}\right)\mathrm{d}t + \mathrm{d}\beta, \qquad s^*(t_{k-1}) = x_{k-1}, \tag{3.218}$$

and the likelihood ratio at t_k can be explicitly written as

$$\begin{aligned} Z(t_k;\omega) &= \frac{\cosh(s^*(t_k))}{\cosh(x_{k-1})} \\ &\times \exp\left[-\left(\frac{m_k - x_{k-1}}{\sqrt{P_k}}\right)(\beta(t_k) - \beta(t_{k-1})) - \frac{1}{2} - \frac{1}{2}\left(\frac{m_k - x_{k-1}}{\sqrt{P_k}}\right)^2\right]. \end{aligned} \tag{3.219}$$

Sequential Importance Resampling Filter for More General SDEs

Next we derive the SIR algorithm for models, where there is an absolutely continuous type of model, which is *embedded* inside a *deterministic* differential equation model. This kind of models are typical in navigation and stochastic control applications, where the deterministic part is typically a plain integral operator. Because the outer operator is deterministic, the likelihood ratios of processes are determined by the inner stochastic processes alone and thus importance sampling of this kind of process is very similar to sampling of the processes considered above.

The following kinds of models are considered:

$$\begin{aligned} \frac{d\mathbf{x}_1}{dt} &= \mathbf{f}_1(\mathbf{x}_1, \mathbf{x}_2, t) & \mathbf{x}_1(0) &= \mathbf{x}_{1,0} \\ d\mathbf{x}_2 &= \mathbf{f}_2(\mathbf{x}_1, \mathbf{x}_2, t)\, dt + \mathbf{L}\, d\boldsymbol{\beta}, & \mathbf{x}_2(0) &= \mathbf{x}_{2,0}, \end{aligned} \tag{3.220}$$

where $\mathbf{f}_1(\cdot)$ and $\mathbf{f}_2(\cdot)$ are deterministic functions, $\boldsymbol{\beta}(t)$ is a Brownian motion and $\mathbf{L}$ is invertible matrix. Note that because the dimensionality of Brownian motion is less than of the joint state $(\mathbf{x}_1\ \mathbf{x}_2)^T$ it is not possible to compute the likelihood ratio between the process and Brownian motion by the Girsanov theorem directly.

However, it turns out that if the importance process for $(\mathbf{x}_1\ \mathbf{x}_2)^T$ is formed as follows

$$\begin{aligned} \frac{d\mathbf{s}_1}{dt} &= \mathbf{f}_1(\mathbf{s}_1, \mathbf{s}_2, t) & \mathbf{s}_1(0) &= \mathbf{x}_{1,0} \\ d\mathbf{s}_2 &= \mathbf{g}_2(\mathbf{s}_1, \mathbf{s}_2, t)\, dt + \mathbf{B}\, d\boldsymbol{\beta}, & \mathbf{s}_2(0) &= \mathbf{x}_{2,0}, \end{aligned} \tag{3.221}$$

then the importance weights can be computed in exactly the same way as when forming importance sample of $\mathbf{x}_2(t)$ using $\mathbf{s}_2(t)$ as the importance process. This is because now $\mathbf{x}_1(t)$ is a *deterministic* functional of $\mathbf{x}_2(t)$ in the sense that given a realization $\{\mathbf{x}_2(t)\ :\ 0 \le t \le T\}$ there is (informally speaking) unique $\{\mathbf{x}_1(t)\ :\ 0 \le t \le T\}$ corresponding to that realization. This deterministic part has no effect on the functional form of the importance weight, provided that it is the same in both the original and importance processes. This can be proved as follows:

Theorem 3.7 (Likelihood ratio of SDEs II)**.** *Assume that processes* $\mathbf{x}_1(t)$*,* $\mathbf{x}_2(t)$*,* $\mathbf{s}_1(t)$ *and* $\mathbf{s}_2(t)$ *are generated by the stochastic differential equations*

$$\frac{d\mathbf{x}_1}{dt} = \mathbf{f}_1(\mathbf{x}_1, \mathbf{x}_2, t), \qquad \mathbf{x}_1(0) = \mathbf{x}_{1,0} \tag{3.222}$$

$$d\mathbf{x}_2 = \mathbf{f}_2(\mathbf{x}_1, \mathbf{x}_2, t)\, dt + \mathbf{L}\, d\boldsymbol{\beta}, \qquad \mathbf{x}_2(0) = \mathbf{x}_{2,0} \tag{3.223}$$

$$\frac{d\mathbf{s}_1}{dt} = \mathbf{f}_1(\mathbf{s}_1, \mathbf{s}_2, t), \qquad \mathbf{s}_1(0) = \mathbf{x}_{1,0} \tag{3.224}$$

$$d\mathbf{s}_2 = \mathbf{g}_2(\mathbf{s}_1, \mathbf{s}_2, t)\, dt + \mathbf{B}\, d\boldsymbol{\beta}, \qquad \mathbf{s}_2(0) = \mathbf{x}_{2,0}, \tag{3.225}$$

where $\mathbf{f}_1$, $\mathbf{f}_2$ *and* $\mathbf{g}_2$ *are bounded and measurable,* $\mathbf{L}$ *and* $\mathbf{B}$ *are invertible matrices and* $\boldsymbol{\beta}(t)$ *is a Brownian motion with respect to measure* P*. Then the expectations of* $\mathbf{x}(t)$ *under measure* P *can be expressed as*

$$\mathrm{E}[\mathbf{h}(\mathbf{x}_1(t), \mathbf{x}_2(t))] = \mathrm{E}[Z(t;\omega)\,\mathbf{h}(\mathbf{s}_1^*(t), \mathbf{s}_2^*(t))], \tag{3.226}$$

where the equations for the scaled processes $\mathbf{s}_1^*(t)$ *and* $\mathbf{s}_2^*(t)$ *are*

$$\frac{\mathrm{d}\mathbf{s}_1^*}{\mathrm{d}t} = \mathbf{f}_1(\mathbf{s}_1^*, \mathbf{s}_2^*, t), \qquad \mathbf{s}_1^*(0) = \mathbf{x}_{1,0} \tag{3.227}$$

$$\mathbf{s}_2^*(t) = \mathbf{x}_{2,0} + \mathbf{L}\,\mathbf{B}^{-1}\,(\mathbf{s}(t) - \mathbf{x}_{2,0}), \tag{3.228}$$

and the likelihood ratio $(\mathrm{d}P_x/\mathrm{d}P_{s*})(t;\omega) = Z(t;\omega)$ *is*

$$\begin{aligned} Z(t;\omega) = \exp\Bigg(& \int_0^t \left[\mathbf{L}^{-1}\,\mathbf{f}_2(\mathbf{s}_1^*(t), \mathbf{s}_2^*(t), t) - \mathbf{B}^{-1}\,\mathbf{g}_2(\mathbf{s}_1(t), \mathbf{s}_2(t), t)\right]^T \,\mathrm{d}\boldsymbol{\beta}(t) \\ & + \int_0^t [\mathbf{L}^{-1}\,\mathbf{f}_2(\mathbf{s}_1^*(t), \mathbf{s}_2^*(t), t)]^T\,[\mathbf{B}^{-1}\,\mathbf{g}_2(\mathbf{s}_1(t), \mathbf{s}_2(t), t)]\,\mathrm{d}t \\ & - \frac{1}{2}\int_0^t \left(||\mathbf{L}^{-1}\mathbf{f}_2(\mathbf{s}_1^*(t), \mathbf{s}_2^*(t), t)||^2 + ||\mathbf{B}^{-1}\mathbf{g}_2(\mathbf{s}_1(t), \mathbf{s}_2(t), t)||^2\right)\mathrm{d}t\Bigg). \end{aligned} \tag{3.229}$$

Proof. As in the proof of Theorem 3.6 define

$$\begin{aligned} Z'(t;\omega) = \exp\Bigg(& -\int_0^t \left[\mathbf{B}^{-1}\,\mathbf{g}_2(\mathbf{s}_1(t), \mathbf{s}_2(t), t)\right]^T \,\mathrm{d}\boldsymbol{\beta}(t;\omega) \\ & - \frac{1}{2}\int_0^t ||\mathbf{B}^{-1}\,\mathbf{g}_2(\mathbf{s}_1(t), \mathbf{s}_2(t), t)||^2\,\mathrm{d}t\Bigg), \end{aligned} \tag{3.230}$$

then because processes $\mathbf{s}_1(t)$ and $\mathbf{s}_2(t)$ are both adapted to the Brownian motion $\boldsymbol{\beta}(t)$, by Theorem 2.5 $\mathbf{n}(t) = \mathbf{B}^{-1}(\mathbf{s}_2(t) - \mathbf{x}_{2,0})$ is a standard Brownian motion with respect to measure $P'(\mathrm{d}\omega) = Z'(t;\omega)\,P(\mathrm{d}\omega)$. Now define the process $Z''(t;\omega)$ analogously to the Equation (3.200):

$$\begin{aligned} Z''(t;\omega) = \exp\Bigg(& \int_0^t \left[\mathbf{L}^{-1}\,\mathbf{f}_2(\mathbf{x}_0 + \mathbf{L}\,\mathbf{n}(t;\omega), t)\right]^T \,\mathrm{d}\mathbf{n}(t;\omega) \\ & - \frac{1}{2}\int_0^t ||\mathbf{L}^{-1}\,\mathbf{f}_2(\mathbf{x}_0 + \mathbf{L}\,\mathbf{n}(t;\omega), t)||^2\,\mathrm{d}t\Bigg) \end{aligned} \tag{3.231}$$

The expectation of any deterministic functional $\mathcal{H}(\mathbf{x}_2(t))$, which depends only on the past $\{\mathbf{x}_2(\tau) \;:\; 0 \le \tau \le t\}$, with respect to the measure $P'(\mathrm{d}\omega)$ can be now written as

$$\begin{aligned}\mathrm{E}'[\mathcal{H}(\mathbf{x}_2(t))] &= \mathrm{E}'[Z''(t;\omega)\,\mathcal{H}(\mathbf{x}_0 + \mathbf{L}\,\mathbf{n}(t))] \\ &= \mathrm{E}'[Z''(t;\omega)\,\mathcal{H}(\mathbf{s}_2^*(t))],\end{aligned} \tag{3.232}$$

where

$$\mathbf{s}_2^*(t) = \mathbf{x}_0 + \mathbf{L}\,\mathbf{B}^{-1}(\mathbf{s}_2(t) - \mathbf{x}_{2,0}). \tag{3.233}$$

Taking the functional to be $\mathcal{H} : \mathbf{x}_2 \mapsto \mathbf{h}(\mathbf{x}_1, \mathbf{x}_2)$, where

$$\frac{\mathrm{d}\mathbf{x}_1}{\mathrm{d}t} = \mathbf{f}_1(\mathbf{x}_1, \mathbf{x}_2, t), \qquad \mathbf{x}_1(0) = \mathbf{x}_{1,0} \tag{3.234}$$

gives that the expectation of an arbitrary function $\mathbf{h}(\mathbf{x}_1, \mathbf{x}_2)$ can be expressed as

$$\mathrm{E}'[\mathbf{h}(\mathbf{x}_1(t), \mathbf{x}_2(t))] = \mathrm{E}'[Z''(t;\omega)\,\mathbf{h}(\mathbf{s}_1^*(t), \mathbf{s}_2^*(t))]. \tag{3.235}$$

where the equations for the scaled processes can be written as (3.227) and (3.228). With respect to the measure P the expectation can be now written as

$$\mathrm{E}[\mathbf{h}(\mathbf{x}_1(t), \mathbf{x}_2(t))] = \mathrm{E}[Z(t;\omega)\,\mathbf{h}(\mathbf{s}_1^*(t), \mathbf{s}_2^*(t))], \tag{3.236}$$

where the explicit expression for $Z(t;\omega) = Z'(t;\omega)\,Z''(t;\omega)$ can be derived in the same way as in the proof of Theorem 3.6. □

Algorithm 3.29 (Importance sampling of SDE II)**.** *Weighted sample from the process $(\mathbf{x}_1(t)\ \mathbf{x}_2(t))^T$ at time T, using the process $(\mathbf{s}_1(t)\ \mathbf{s}_2(t))^T$ as the importance process, can be generated as follows:*

1. *Draw N Brownian motions $\{\boldsymbol{\beta}^{(i)}(t), 0 \le t \le T, i = 1, \ldots, N\}$ and simulate the importance process*

$$\frac{\mathrm{d}\mathbf{s}_1^{(i)}}{\mathrm{d}t} = \mathbf{f}_1(\mathbf{s}_1^{(i)}, \mathbf{s}_2^{(i)}, t), \qquad \mathbf{s}_1^{(i)}(0) = \mathbf{x}_{1,0} \tag{3.237}$$

$$\mathrm{d}\mathbf{s}_2^{(i)} = \mathbf{g}_2(\mathbf{s}_1^{(i)}, \mathbf{s}_2^{(i)}, t)\,\mathrm{d}t + \mathbf{B}\,\mathrm{d}\boldsymbol{\beta}^{(i)}, \qquad \mathbf{s}_2^{(i)}(0) = \mathbf{x}_{2,0}, \tag{3.238}$$

from $t = 0$ to $t = T$, compute the scaled importance processes

$$\frac{\mathrm{d}\mathbf{s}_1^{*(i)}}{\mathrm{d}t} = \mathbf{f}_1(\mathbf{s}_1^{*(i)}, \mathbf{s}_2^{*(i)}, t), \qquad \mathbf{s}_1^{*(i)}(0) = \mathbf{x}_{1,0} \tag{3.239}$$

$$\mathbf{s}_2^{*(i)}(t) = \mathbf{x}_{2,0} + \mathbf{L}\,\mathbf{B}^{-1}\,(\mathbf{s}_2^{(i)}(t) - \mathbf{x}_{2,0}), \tag{3.240}$$

and set

$$\tilde{\mathbf{x}}_1^{(i)} = \mathbf{s}_1^{*(i)}(T) \tag{3.241}$$

$$\tilde{\mathbf{x}}_2^{(i)} = \mathbf{s}_2^{*(i)}(T). \tag{3.242}$$

2. *For each i compute*

$$\begin{aligned}
w^{(i)} = \exp\Bigg(\int_0^T & \Big[\mathbf{L}^{-1}\,\mathbf{f}_2(\mathbf{s}_1^{*(i)}(t),\mathbf{s}_2^{*(i)}(t),t) \\
& - \mathbf{B}^{-1}\,\mathbf{g}_2(\mathbf{s}_1^{(i)}(t),\mathbf{s}_2^{(i)}(t),t)\Big]^T \,\mathrm{d}\boldsymbol{\beta}^{(i)} \\
& + \int_0^T [\mathbf{L}^{-1}\,\mathbf{f}_2(\mathbf{s}_1^{*(i)}(t),\mathbf{s}_2^{*(i)}(t),t)]^T\,[\mathbf{B}^{-1}\,\mathbf{g}_2(\mathbf{s}_1^{(i)}(t),\mathbf{s}_2^{(i)}(t),t)]\,\mathrm{d}t \\
& - \frac{1}{2}\int_0^T \Big(||\mathbf{L}^{-1}\,\mathbf{f}_2(\mathbf{s}_1^{*(i)}(t),\mathbf{s}_2^{*(i)}(t),t)||^2 \\
& + ||\mathbf{B}^{-1}\,\mathbf{g}_2(\mathbf{s}_1^{(i)}(t),\mathbf{s}_2^{(i)}(t),t)||^2\Big)\,\mathrm{d}t\Bigg).
\end{aligned} \tag{3.243}$$

3. *Now $\{(\tilde{\mathbf{x}}_1^{(i)},\tilde{\mathbf{x}}_2^{(i)},w^{(i)}) \;:\; i=1,\ldots,N\}$ is a set of importance samples such that for any function $\mathbf{h}(\cdot)$*

$$\mathrm{E}[\mathbf{h}(\mathbf{x}_1(T),\mathbf{x}_2(T))] \approx \sum_i w^{(i)}\,\mathbf{h}(\tilde{\mathbf{x}}_1^{(i)},\tilde{\mathbf{x}}_2^{(i)}), \tag{3.244}$$

where $(\mathbf{x}_1(T),\mathbf{x}_2(T))^T$ is the solution to the stochastic differential equation

$$\frac{\mathrm{d}\mathbf{x}_1}{\mathrm{d}t} = \mathbf{f}_1(\mathbf{x}_1,\mathbf{x}_2,t) \qquad \mathbf{x}_1(0) = \mathbf{x}_{1,0} \tag{3.245}$$

$$\mathrm{d}\mathbf{x}_2 = \mathbf{f}_2(\mathbf{x}_1,\mathbf{x}_2,t)\,\mathrm{d}t + \mathbf{L}\,\mathrm{d}\boldsymbol{\beta}, \qquad \mathbf{x}_2(0) = \mathbf{x}_{2,0}, \tag{3.246}$$

at time T.

The continuous-discrete SIR Algorithm 3.28 can be now generalized to filtering models of the form

$$\begin{aligned}
\frac{\mathrm{d}\mathbf{x}_1}{\mathrm{d}t} &= \mathbf{f}_1(\mathbf{x}_1,\mathbf{x}_2,t) \\
\mathrm{d}\mathbf{x}_2 &= \mathbf{f}_2(\mathbf{x}_1,\mathbf{x}_2,t)\,\mathrm{d}t + \mathbf{L}\,\mathrm{d}\boldsymbol{\beta} \\
\mathbf{y}_k &\sim p(\mathbf{y}_k\,|\,\mathbf{x}(t_k)),
\end{aligned} \tag{3.247}$$

where the state consists of two components $\mathbf{x}(t) = (\mathbf{x}_1(t),\mathbf{x}_2(t))$, $\mathbf{f}_1(\cdot)$ and $\mathbf{f}_2(\cdot)$ are deterministic functions, $\boldsymbol{\beta}(t)$ is a Brownian motion and $\mathbf{L}$ is invertible matrix. When the importance process is selected to be of the form

$$\begin{aligned}
\frac{\mathrm{d}\mathbf{s}_1}{\mathrm{d}t} &= \mathbf{f}_1(\mathbf{s}_1,\mathbf{s}_2,t) \\
\mathrm{d}\mathbf{s}_2 &= \mathbf{g}_2(\mathbf{s}_1,\mathbf{s}_2,t)\,\mathrm{d}t + \mathbf{B}\,\mathrm{d}\boldsymbol{\beta},
\end{aligned} \tag{3.248}$$

then importance sampling can be performed with the Algorithm 3.29. The SIR algorithm takes now the following form:

Algorithm 3.30 (Continuous-discrete SIR II)**.** *Given the importance process, a weighted set of samples* $\{\tilde{\mathbf{x}}^{(i)}_{1,k-1}, \tilde{\mathbf{x}}^{(i)}_{2,k-1}, w^{(i)}_{k-1}\}$ *and the new measurement* $\mathbf{y}_k$*, a single step of continuous-discrete sequential importance resampling can be now performed as follows:*

1. *Draw* N *Brownian motions* $\{\boldsymbol{\beta}^{(i)}(t), t_{k-1} \le t \le t_k, i = 1, \ldots, N\}$ *and simulate the importance process*

$$\frac{\mathrm{d}\mathbf{s}^{(i)}_1}{\mathrm{d}t} = \mathbf{f}_1(\mathbf{s}^{(i)}_1, \mathbf{s}^{(i)}_2, t), \qquad \mathbf{s}^{(i)}_1(t_{k-1}) = \tilde{\mathbf{x}}^{(i)}_{1,k-1} \tag{3.249}$$

$$\mathrm{d}\mathbf{s}^{(i)}_2 = \mathbf{g}_2(\mathbf{s}^{(i)}_1, \mathbf{s}^{(i)}_2, t)\,\mathrm{d}t + \mathbf{B}\,\mathrm{d}\boldsymbol{\beta}^{(i)}, \qquad \mathbf{s}^{(i)}_2(t_{k-1}) = \tilde{\mathbf{x}}^{(i)}_{2,k-1}, \tag{3.250}$$

from $t = t_{k-1}$ *to* $t = t_k$*, and compute*

$$\frac{\mathrm{d}\mathbf{s}^{*(i)}_1}{\mathrm{d}t} = \mathbf{f}_1(\mathbf{s}^{*(i)}_1, \mathbf{s}^{*(i)}_2, t), \qquad \mathbf{s}^{*(i)}_1(0) = \mathbf{x}^{(i)}_{1,k-1} \tag{3.251}$$

$$\mathbf{s}^{*(i)}_2(t) = \mathbf{x}^{(i)}_{2,k-1} + \mathbf{L}\,\mathbf{B}^{-1}\,(\mathbf{s}^{(i)}_2(t) - \mathbf{x}^{(i)}_{2,k-1}), \tag{3.252}$$

and set

$$\tilde{\mathbf{x}}^{(i)}_{1,k} = \mathbf{s}^{*(i)}_1(t_k) \tag{3.253}$$

$$\tilde{\mathbf{x}}^{(i)}_{2,k} = \mathbf{s}^{*(i)}_2(t_k). \tag{3.254}$$

2. *For each* i *compute*

$$\begin{aligned} w^{(i)}_k = w^{(i)}_{k-1}\, \exp\Bigg(& \int_{t_{k-1}}^{t_k} \Big[\mathbf{L}^{-1}\,\mathbf{f}_2(\mathbf{s}^{*(i)}_1(t), \mathbf{s}^{*(i)}_2(t), t) \\ & \quad - \mathbf{B}^{-1}\,\mathbf{g}_2(\mathbf{s}^{(i)}_1(t), \mathbf{s}^{(i)}_2(t), t)\Big]^T \mathrm{d}\boldsymbol{\beta}^{(i)} \\ & + \int_{t_{k-1}}^{t_k} [\mathbf{L}^{-1}\,\mathbf{f}_2(\mathbf{s}^{*(i)}_1(t), \mathbf{s}^{*(i)}_2(t), t)]^T\,[\mathbf{B}^{-1}\,\mathbf{g}_2(\mathbf{s}^{(i)}_1(t), \mathbf{s}^{(i)}_2(t), t)]\,\mathrm{d}t \\ & - \frac{1}{2}\int_{t_{k-1}}^{t_k} \Big(||\mathbf{L}^{-1}\,\mathbf{f}_2(\mathbf{s}^{*(i)}_1(t), \mathbf{s}^{*(i)}_2(t), t)||^2 \\ & \quad + ||\mathbf{B}^{-1}\,\mathbf{g}_2(\mathbf{s}^{(i)}_1(t), \mathbf{s}^{(i)}_2(t), t)||^2\Big)\,\mathrm{d}t\Bigg) \\ & \quad \times p(\mathbf{y}_k \,|\, \tilde{\mathbf{x}}^{(i)}_{1,k}, \tilde{\mathbf{x}}^{(i)}_{2,k}), \end{aligned} \tag{3.255}$$

and re-normalize the weights to sum to unity.

3. *If the effective number of weights* (3.97) *is too low, perform resampling.*

Analogously to the discrete-time sequential importance sampling it would be now desirable to use (optimal) importance process, which would have the law

$$\begin{aligned}&\pi(\mathbf{x}(t) \,|\, \mathbf{y}_{1:k}, \tilde{\mathbf{x}}^{(i)}_{k-1}) \\ &\quad = p(\mathbf{x}(t) \,|\, \mathbf{y}_k, \mathbf{x}(t_{k-1}) = \tilde{\mathbf{x}}^{(i)}_{k-1}), \qquad t_k < t < t_{k-1},\end{aligned} \tag{3.256}$$

where $\mathbf{y}_k$ is the (new) measurement at time instance t_k. This solution is the continuous-time smoothing solution, when the smoother is started from initial conditions $\mathbf{x}(t_{k-1}) = \tilde{\mathbf{x}}^{(i)}_{k-1}$ and (the filter and) the smoother processes only the single measurement $\mathbf{y}_k$. However, in order to apply the importance sampling theory presented here the importance process should be a Markov process.

As in the discrete-time case it could be possible to use continuous-discrete versions of the extended Kalman filters/smoothers and unscented Kalman filters/smoothers for forming the importance processes. However, further investigation of the possibility of using the smoothers is left to the future.

Example 3.12 (CD-SIR estimation of noisy physical pendulum)**.** *The stochastic differential equation for the angular position of a physical pendulum (Alonso and Finn, 1980), which is distorted by random white noise accelerations $w(t)$ with spectral density q can be written as*

$$\frac{\mathrm{d}^2 x}{\mathrm{d}t^2} + a^2 \, \sin(x) = w(t). \tag{3.257}$$

where a is the angular velocity of the (linearized) pendulum. If we define the state as $\mathbf{x} = (x \; \mathrm{d}x/\mathrm{d}t)^T$ and change to state space form and to the integral equation notation in terms of standard Brownian motion the model can be written as

$$\begin{aligned}\frac{\mathrm{d}x_1}{\mathrm{d}t} &= x_2 \\ \mathrm{d}x_2 &= -a^2 \, \sin(x_1) \, \mathrm{d}t + q^{1/2} \, \mathrm{d}\beta,\end{aligned} \tag{3.258}$$

which is model of the form (3.220).

Assuming that the measurement model is such form that it can be approximated by a Gaussian model, an importance process can be now formed by using either EKF or UKF using a Gaussian approximation to the measurement model and the result is a 2-dimensional Gaussian approximation for the joint distribution of the state $\mathbf{x}(t_k) = (x_1(t_k) \; x_2(t_k))^T$. If the marginal mean and covariance of $x_2(t_k)$ are $m_{2,k}$ and $P_{22,k}$, then a suitable importance process is (assuming that *sampling interval is Δt)*

$$\begin{aligned}
\frac{\mathrm{d}s_1}{\mathrm{d}t} &= s_2 & s_1(t_{k-1}) &= x_{1,k-1} \\
\mathrm{d}s_2 &= \left(\frac{m_{2,k} - x_{2,k-1}}{\Delta t}\right) \mathrm{d}t + \sqrt{\frac{P_{22,k}}{\Delta t}}\, \mathrm{d}\beta & s_2(t_{k-1}) &= x_{2,k-1}.
\end{aligned} \tag{3.259}$$

The equations for the scaled importance process can be now written as

$$\begin{aligned}
\frac{\mathrm{d}s_1^*}{\mathrm{d}t} &= s_2^*, & s_1^*(t_{k-1}) &= x_{1,k-1} \\
\mathrm{d}s_2^* &= \left(\sqrt{\frac{q}{P_{22,k}\,\Delta t}}\right) (m_{2,k} - x_{2,k-1})\, \mathrm{d}t + q^{1/2}\, \mathrm{d}\beta, & s_2^*(t_{k-1}) &= x_{2,k-1},
\end{aligned} \tag{3.260}$$

and the likelihood ratio is given by the formula

$$\begin{aligned}
Z(t;\omega) = \exp\Bigg[&\int -\frac{a^2}{q^{1/2}} \sin(s_1^*(t))\, \mathrm{d}\beta - \int \left(\frac{m_{2,k} - x_{2,k-1}}{\sqrt{P_{22,k}\,\Delta t}}\right) \mathrm{d}\beta \\
&- \int \frac{a^2}{q^{1/2}} \sin(s_1^*(t)) \left(\frac{m_{2,k} - x_{2,k-1}}{\sqrt{P_{22,k}\,\Delta t}}\right) \mathrm{d}t \\
&- \frac{1}{2} \int \frac{a^4}{q} \sin^2(s_1^*(t))\, \mathrm{d}t - \frac{1}{2} \int \left(\frac{m_{2,k} - x_{2,k-1}}{\sqrt{P_{22,k}\,\Delta t}}\right)^2 \mathrm{d}t \Bigg].
\end{aligned} \tag{3.261}$$

Continuous-Discrete Sequential Importance Resampling Smoother

As in the case of discrete-time SIR smoothing in Section 3.1.3, *optimal continuous-discrete smoothing* can be performed with the continuous-discrete SIR by keeping the whole trajectories instead of the samples at measurement times. However, this approximation can be quite degenerate and more efficient methods using the ideas in (Godsill et al., 2004) could be developed.

Rao-Blackwellized Sequential Importance Resampling Filter

Next the SIR algorithm for the class of models is derived, where a *linear* stochastic differential equation is driven by a model, from which it is possible to generate importance samples by the methods already described in this Section. This kind of models can be handled such that only the inner process is sampled and the linear part is integrated out using the continuous-discrete Kalman filter. Then it is possible to form a Rao-Blackwellized estimate, where the probability density is approximated by a mixture of Gaussian distributions. The measurement model is assumed to be of the same form as in previous sections, but linear with respect to the state variables corresponding to the linear part of the dynamic process.

Consider a model of the form

$$\begin{aligned} \mathrm{d}\mathbf{x}_1 &= \mathbf{F}(\mathbf{x}_2,\mathbf{x}_3,t)\,\mathbf{x}_1\,\mathrm{d}t + \mathbf{f}_1(\mathbf{x}_2,\mathbf{x}_3,t)\,\mathrm{d}t \\ &\quad + \mathbf{V}(\mathbf{x}_2,\mathbf{x}_3,t)\,\mathrm{d}\boldsymbol{\eta}, && \mathbf{x}_1(0) \sim \mathrm{N}(\mathbf{m}_0,\mathbf{P}_0) \\ \frac{\mathrm{d}\mathbf{x}_2}{\mathrm{d}t} &= \mathbf{f}_2(\mathbf{x}_2,\mathbf{x}_3,t) && \mathbf{x}_2(0) = \mathbf{x}_{2,0} \\ \mathrm{d}\mathbf{x}_3 &= \mathbf{f}_3(\mathbf{x}_2,\mathbf{x}_3,t)\,\mathrm{d}t + \mathbf{L}\,\mathrm{d}\boldsymbol{\beta}, && \mathbf{x}_3(0) = \mathbf{x}_{3,0}, \end{aligned} \tag{3.262}$$

where $\boldsymbol{\beta}(t)$ and $\boldsymbol{\eta}(t)$ are independent Brownian motions. In this case an importance process can be formed as

$$\begin{aligned} \mathrm{d}\mathbf{s}_1 &= \mathbf{F}(\mathbf{s}_2,\mathbf{s}_3,t)\,\mathbf{s}_1\,\mathrm{d}t + \mathbf{f}_1(\mathbf{s}_2,\mathbf{s}_3,t)\,\mathrm{d}t \\ &\quad + \mathbf{V}(\mathbf{s}_2,\mathbf{s}_3,t)\,\mathrm{d}\boldsymbol{\eta}, && \mathbf{s}_1(0) \sim \mathrm{N}(\mathbf{m}_0,\mathbf{P}_0) \\ \frac{\mathrm{d}\mathbf{s}_2}{\mathrm{d}t} &= \mathbf{f}_2(\mathbf{s}_2,\mathbf{s}_3,t) && \mathbf{s}_2(0) = \mathbf{x}_{2,0} \\ \mathrm{d}\mathbf{s}_3 &= \mathbf{g}_3(\mathbf{s}_2,\mathbf{s}_3,t)\,\mathrm{d}t + \mathbf{B}\,\mathrm{d}\boldsymbol{\beta}, && \mathbf{s}_3(0) = \mathbf{x}_{3,0}, \end{aligned} \tag{3.263}$$

In both the original and importance processes, conditionally to the filtration of the second Brownian motion $\mathscr{F}_t = \sigma(\boldsymbol{\beta}(s), 0 \le s \le t)$ the law of the first equation is determined by the mean and covariance of the Gaussian process, which is driven by the process $\boldsymbol{\eta}(t)$. Thus, conditionally to the filtration $\mathscr{F}_t$ of $\boldsymbol{\beta}(t)$ (and thus $\mathbf{x}_2$ and $\mathbf{x}_3$) the process $\mathbf{x}_1(t)$ is Gaussian for all t. The same applies to the importance process.

Now it is possible to integrate out the Gaussian parts of both the processes. This procedure results in the following marginalized equations for the original process:

$$\begin{aligned} \frac{\mathrm{d}\mathbf{m}_x(t)}{\mathrm{d}t} &= \mathbf{F}(\mathbf{x}_2,\mathbf{x}_3,t)\,\mathbf{m}_x(t) + \mathbf{f}_1(\mathbf{x}_2,\mathbf{x}_3,t), && \mathbf{m}_x(0) = \mathbf{m}_0, \\ \frac{\mathrm{d}\mathbf{P}_x(t)}{\mathrm{d}t} &= \mathbf{F}(\mathbf{x}_2,\mathbf{x}_3,t)\,\mathbf{P}_x(t) + \mathbf{P}_x(t)\,\mathbf{F}^T(\mathbf{x}_2,\mathbf{x}_3,t) \\ &\quad + \mathbf{V}(\mathbf{x}_2,\mathbf{x}_3,t)\,\mathbf{V}^T(\mathbf{x}_2,\mathbf{x}_3,t), && \mathbf{P}_x(0) = \mathbf{P}_0 \\ \frac{\mathrm{d}\mathbf{x}_2}{\mathrm{d}t} &= \mathbf{f}_2(\mathbf{x}_2,\mathbf{x}_3,t) && \mathbf{x}_2(0) = \mathbf{x}_{2,0} \\ \mathrm{d}\mathbf{x}_3 &= \mathbf{f}_3(\mathbf{x}_2,\mathbf{x}_3,t)\,\mathrm{d}t + \mathbf{L}\,\mathrm{d}\boldsymbol{\beta}, && \mathbf{x}_3(0) = \mathbf{x}_{3,0}, \end{aligned} \tag{3.264}$$

where $\mathbf{m}_x(t)$ and $\mathbf{P}_x(t)$ are the mean and covariance of the Gaussian process. For the importance process we get similarly:

$$\begin{aligned}
\frac{\mathrm{d}\mathbf{m}_s(t)}{\mathrm{d}t} &= \mathbf{F}(\mathbf{s}_2, \mathbf{s}_3, t)\, \mathbf{m}_s(t) + \mathbf{f}_1(\mathbf{s}_2, \mathbf{s}_3, t), & \mathbf{m}_s(0) &= \mathbf{m}_0, \\
\frac{\mathrm{d}\mathbf{P}_s(t)}{\mathrm{d}t} &= \mathbf{F}(\mathbf{s}_2, \mathbf{s}_3, t)\, \mathbf{P}_s(t) + \mathbf{P}_s(t)\, \mathbf{F}^T(\mathbf{s}_2, \mathbf{s}_3, t) \\
&\quad + \mathbf{V}(\mathbf{s}_2, \mathbf{s}_3, t)\, \mathbf{V}^T(\mathbf{s}_2, \mathbf{s}_3, t), & \mathbf{P}_s(0) &= \mathbf{P}_0 \\
\frac{\mathrm{d}\mathbf{s}_2}{\mathrm{d}t} &= \mathbf{f}_2(\mathbf{s}_2, \mathbf{s}_3, t) & \mathbf{s}_2(0) &= \mathbf{x}_{2,0} \\
\mathrm{d}\mathbf{s}_3 &= \mathbf{g}_3(\mathbf{s}_2, \mathbf{s}_3, t)\, \mathrm{d}t + \mathbf{B}\, \mathrm{d}\boldsymbol{\beta}, & \mathbf{s}_3(0) &= \mathbf{x}_{3,0},
\end{aligned} \tag{3.265}$$

The models (3.264) and (3.265) have now the form, where the Algorithm 3.29 can be used. The importance sampling now results in the set of weighted samples

$$\{w^{(i)}, \tilde{\mathbf{m}}^{(i)}, \tilde{\mathbf{P}}^{(i)}, \tilde{\mathbf{x}}_2^{(i)}, \tilde{\mathbf{x}}_3^{(i)}\} \tag{3.266}$$

such that the probability density of the state $\mathbf{x}(T) = (\mathbf{x}_1(T), \mathbf{x}_2(T), \mathbf{x}_3(T))$ at time T is approximately given as

$$\begin{aligned}
&p(\mathbf{x}_1(T), \mathbf{x}_2(T), \mathbf{x}_3(T)) \\
&\approx \sum_i w^{(i)}\ \mathrm{N}(\mathbf{x}_1(T) \mid \tilde{\mathbf{m}}^{(i)}, \tilde{\mathbf{P}}^{(i)})\, \delta(\mathbf{x}_2(T) - \tilde{\mathbf{x}}_2^{(i)})\, \delta(\mathbf{x}_3(T) - \tilde{\mathbf{x}}_3^{(i)}).
\end{aligned} \tag{3.267}$$

If the measurement model is of the form

$$\mathbf{y}_k = \mathbf{H}_k\left(\mathbf{x}_2(t_k), \mathbf{x}_3(t_k)\right)\ \mathbf{x}_1(t_k) + \mathbf{r}_k, \qquad \mathbf{r}_k \sim \mathrm{N}\left(\mathbf{0}, \mathbf{R}_k\left(\mathbf{x}_2(t_k), \mathbf{x}_3(t_k)\right)\right), \tag{3.268}$$

that is,

$$p(\mathbf{y}_k \mid \mathbf{x}(t_k)) = \mathrm{N}\left(\mathbf{y}_k \mid \mathbf{H}_k\left(\mathbf{x}_2(t_k), \mathbf{x}_3(t_k)\right)\ \mathbf{x}_1(t_k), \mathbf{R}_k\left(\mathbf{x}_2(t_k), \mathbf{x}_3(t_k)\right)\right), \tag{3.269}$$

then conditionally to $\mathbf{x}_2(t_k), \mathbf{x}_3(t_k)$ also the measurement model is linear Gaussian and the Kalman filter update equations can be applied. The resulting algorithm is as follows:

Algorithm 3.31 (Conditionally Gaussian continuous-discrete Rao-Blackwellized SIR)**.** *Given set of importance samples* $\{\tilde{\mathbf{x}}_{2,k-1}^{(i)}, \tilde{\mathbf{x}}_{3,k-1}^{(i)}, \tilde{\mathbf{m}}_{k-1}^{(i)}, \tilde{\mathbf{P}}_{k-1}^{(i)}, w_{k-1}^{(i)} : i = 1, \ldots, N\}$ *and the measurement* $\mathbf{y}_k$ *do the following:*

1. *Draw* N *Brownian motions* $\{\boldsymbol{\beta}^{(i)}(t), t_{k-1} < t \leq t_k, i = 1, \ldots, N\}$ *and*

simulate the importance process

$$\begin{aligned}
\frac{\mathrm{d}\mathbf{m}_s^{(i)}}{\mathrm{d}t} &= \mathbf{F}(\mathbf{s}_2^{(i)}(t), \mathbf{s}_3^{(i)}, t)\, \mathbf{m}_s^{(i)}(t) + \mathbf{f}_1(\mathbf{s}_2^{(i)}, \mathbf{s}_3^{(i)}, t) \\
\frac{\mathrm{d}\mathbf{P}_s^{(i)}}{\mathrm{d}t} &= \mathbf{F}(\mathbf{s}_2^{(i)}, \mathbf{s}_3^{(i)}, t)\, \mathbf{P}_s^{(i)}(t) + \mathbf{P}_s^{(i)}(t)\, \mathbf{F}^T(\mathbf{s}_2^{(i)}, \mathbf{s}_3^{(i)}, t) \\
&\quad + \mathbf{V}(\mathbf{s}_2^{(i)}, \mathbf{s}_3^{(i)}, t)\, \mathbf{V}^T(\mathbf{s}_2^{(i)}, \mathbf{s}_3^{(i)}, t) \\
\frac{\mathrm{d}\mathbf{s}_2^{(i)}}{\mathrm{d}t} &= \mathbf{f}_2(\mathbf{s}_2^{(i)}, \mathbf{s}_3^{(i)}, t) \\
\mathrm{d}\mathbf{s}_3^{(i)} &= \mathbf{g}_3(\mathbf{s}_2^{(i)}, \mathbf{s}_3^{(i)}, t)\, \mathrm{d}t + \mathbf{B}\, \mathrm{d}\boldsymbol{\beta}^{(i)},
\end{aligned} \tag{3.270}$$

with initial conditions

$$\begin{aligned}
\mathbf{m}_s^{(i)}(t_{k-1}) &= \tilde{\mathbf{m}}_{k-1}^{(i)} \\
\mathbf{P}_s^{(i)}(t_{k-1}) &= \tilde{\mathbf{P}}_{k-1}^{(i)} \\
\mathbf{s}_2^{(i)}(t_{k-1}) &= \tilde{\mathbf{x}}_{2,k-1}^{(i)} \\
\mathbf{s}_3^{(i)}(t_{k-1}) &= \tilde{\mathbf{x}}_{3,k-1}^{(i)},
\end{aligned} \tag{3.271}$$

and the scaled importance process

$$\begin{aligned}
\frac{\mathrm{d}\mathbf{m}_s^{*(i)}}{\mathrm{d}t} &= \mathbf{F}(\mathbf{s}_2^{*(i)}(t), \mathbf{s}_3^{*(i)}, t)\, \mathbf{m}_s^{*(i)}(t) + \mathbf{f}_1(\mathbf{s}_2^{*(i)}, \mathbf{s}_3^{*(i)}, t) \\
\frac{\mathrm{d}\mathbf{P}_s^{*(i)}}{\mathrm{d}t} &= \mathbf{F}(\mathbf{s}_2^{*(i)}, \mathbf{s}_3^{*(i)}, t)\, \mathbf{P}_s^{*(i)}(t) + \mathbf{P}_s^{*(i)}(t)\, \mathbf{F}^T(\mathbf{s}_2^{*(i)}, \mathbf{s}_3^{*(i)}, t) \\
&\quad + \mathbf{V}(\mathbf{s}_2^{*(i)}, \mathbf{s}_3^{*(i)}, t)\, \mathbf{V}^T(\mathbf{s}_2^{*(i)}, \mathbf{s}_3^{*(i)}, t) \\
\frac{\mathrm{d}\mathbf{s}_2^{*(i)}}{\mathrm{d}t} &= \mathbf{f}_2(\mathbf{s}_2^{*(i)}, \mathbf{s}_3^{*(i)}, t) \\
\mathbf{s}_3^{*(i)}(t) &= \tilde{\mathbf{x}}_{3,k-1}^{(i)} + \mathbf{L}\, \mathbf{B}^{-1}\, (\mathbf{s}_3^{(i)}(t) - \tilde{\mathbf{x}}_{3,k-1}^{(i)}),
\end{aligned} \tag{3.272}$$

with the same initial conditions from $t = t_{k-1}$ *to* $t = t_k$ *and set*

$$\tilde{\mathbf{m}}_k^{-(i)} = \mathbf{m}_s^{*(i)}(t_k) \tag{3.273}$$

$$\tilde{\mathbf{P}}_k^{-(i)} = \mathbf{P}_s^{*(i)}(t_k) \tag{3.274}$$

$$\tilde{\mathbf{x}}_{2,k}^{(i)} = \mathbf{s}_2^{*(i)}(t_k) \tag{3.275}$$

$$\tilde{\mathbf{x}}_{3,k}^{(i)} = \mathbf{s}_3^{*(i)}(t_k). \tag{3.276}$$

2. *For each i perform the Kalman filter update*

$$
\begin{aligned}
\boldsymbol{\mu}_k^{(i)} &= \mathbf{H}_k(\tilde{\mathbf{x}}_{2,k}^{(i)}, \tilde{\mathbf{x}}_{3,k}^{(i)})\, \mathbf{m}_k^{-(i)} && (3.277)\\
\mathbf{S}_k^{(i)} &= \mathbf{H}_k(\tilde{\mathbf{x}}_{2,k}^{(i)}, \tilde{\mathbf{x}}_{3,k}^{(i)})\, \mathbf{P}_k^{-(i)}\, \mathbf{H}_k^T(\tilde{\mathbf{x}}_{2,k}^{(i)}, \tilde{\mathbf{x}}_{3,k}^{(i)}) + \mathbf{R}_k(\tilde{\mathbf{x}}_{2,k}^{(i)}, \tilde{\mathbf{x}}_{3,k}^{(i)}) && (3.278)\\
\mathbf{K}_k^{(i)} &= \mathbf{P}_k^{-(i)}\, \mathbf{H}_k^T(\tilde{\mathbf{x}}_{2,k}^{(i)}, \tilde{\mathbf{x}}_{3,k}^{(i)})\, [\mathbf{S}_k^{(i)}]^{-1} && (3.279)\\
\mathbf{m}_k^{(i)} &= \mathbf{m}_k^{-(i)} + \mathbf{K}_k^{(i)}\, (\mathbf{y}_k - \boldsymbol{\mu}_k^{(i)}) && (3.280)\\
\mathbf{P}_k^{(i)} &= \mathbf{P}_k^{-(i)} - \mathbf{K}_k^{(i)}\, \mathbf{S}_k^{(i)}\, [\mathbf{K}_k^{(i)}]^T, && (3.281)
\end{aligned}
$$

and compute the importance weight

$$
\begin{aligned}
w_k^{(i)} &= w_{k-1}^{(i)} \exp\Bigg(\int_{t_{k-1}}^{t_k} \Big[\mathbf{L}^{-1}\, \mathbf{f}_3(\mathbf{s}_2^{*(i)}(t), \mathbf{s}_3^{*(i)}(t), t) \\
&\qquad - \mathbf{B}^{-1}\, \mathbf{g}_3(\mathbf{s}_2^{(i)}(t), \mathbf{s}_3^{(i)}(t), t)\Big]^T \, \mathrm{d}\boldsymbol{\beta}^{(i)} \\
&\quad + \int_{t_{k-1}}^{t_k} [\mathbf{L}^{-1}\, \mathbf{f}_3(\mathbf{s}_2^{*(i)}(t), \mathbf{s}_3^{*(i)}(t), t)]^T\, [\mathbf{B}^{-1}\, \mathbf{g}_3(\mathbf{s}_2^{(i)}(t), \mathbf{s}_3^{(i)}(t), t)]\, \mathrm{d}t \\
&\quad - \frac{1}{2} \int_{t_{k-1}}^{t_k} \Big(||\mathbf{L}^{-1}\, \mathbf{f}_3(\mathbf{s}_2^{*(i)}(t), \mathbf{s}_3^{*(i)}(t), t)||^2 \\
&\qquad + ||\mathbf{B}^{-1}\, \mathbf{g}_3(\mathbf{s}_2^{(i)}(t), \mathbf{s}_3^{(i)}(t), t)||^2 \Big)\, \mathrm{d}t \Bigg) \\
&\qquad \times \mathrm{N}(\mathbf{y}_k \mid \boldsymbol{\mu}_k^{(i)}, \mathbf{S}_k^{(i)}),
\end{aligned}
\tag{3.282}
$$

and re-normalize the weights to sum to unity.

3. *If the effective number of weights* (3.97) *is too low, perform resampling.*

Example 3.13 (Noisy pendulum problem with correlated noise)**.** *Consider again the noisy pendulum in Example 3.12. Assume that the measurement noise is autocorrelated with approximately the following spectral density*

$$
S(\omega) = \frac{1}{\gamma^2 + \omega^2}. \tag{3.283}
$$

The stochastic differential equation of the measurement noise $v(t)$ is thus given as

$$
\mathrm{d}v = -\gamma\, v\, \mathrm{d}t + \mathrm{d}\eta \tag{3.284}
$$

where η is a standard Brownian motion. The measurement y_k is the state of the pendulum plus the correlated noise $v(t_k)$:

$$
y_k = x_1(t_k) + v(t_k). \tag{3.285}
$$

Further assume that the correlation parameter γ actually depends on the state of the pendulum as follows:

$$\gamma(x_1) = c\,x_1^2. \tag{3.286}$$

That is, the correlation depends linearly on the distance of the pendulum from the center position. The whole model can be now written as

$$\begin{aligned}
\mathrm{d}v &= -c\,x_1^2\,v\,\mathrm{d}t + \mathrm{d}\eta \\
\frac{\mathrm{d}x_1}{\mathrm{d}t} &= x_2 \\
\mathrm{d}x_2 &= -a^2\,\sin(x_1)\,\mathrm{d}t + q^{1/2}\,\mathrm{d}\beta \\
y_k &= x_1(t_k) + v(t_k),
\end{aligned} \tag{3.287}$$

which is a model suitable to the continuous-discrete Rao-Blackwellized particle filter presented in this section. Note that because there is no white noise term in the measurement likelihood, a plain particle filter could not be applied to this problem at all, except by including artificial white measurement noise. However, extended and unscented Kalman filter can be applied to the problem as such.

An importance process for the model can be formed by computing a Gaussian approximation by EKF or UKF and retaining only the part corresponding to the process x_2. The importance process can be then formed in the same manner as in Example 3.12 and the likelihood ratio can be computed in exactly the same way.

Sometimes, when the model is not exactly conditionally linear, it is still possible to approximately marginalize some of the state components by using continuous-discrete extended or unscented Kalman filters.

Analogously to the discrete-time case (Storvik, 2002) Rao-Blackwellization can often be applied to models with unknown static parameters having the general form

$$\begin{aligned}
\mathrm{d}\mathbf{x} &= \mathbf{f}(\mathbf{x},\boldsymbol{\theta},t)\,\mathrm{d}t + \mathbf{L}(\mathbf{x},\boldsymbol{\theta},t)\,\mathrm{d}\boldsymbol{\beta} \\
\mathbf{y}_k &\sim p(\mathbf{y}_k \,|\, \mathbf{x}(t_k),\boldsymbol{\theta}) \\
\boldsymbol{\theta} &\sim p(\boldsymbol{\theta}),
\end{aligned} \tag{3.288}$$

where vector $\boldsymbol{\theta}$ contains the unknown static parameters. If the posterior distribution of the unknown static parameters $\boldsymbol{\theta}$ depends only on a suitable set of sufficient statistics $\mathbf{T}_k = \mathbf{T}_k(\mathbf{x}_{1:k},\mathbf{y}_{1:k})$, the parameter can be marginalized out analytically and only the state needs to be sampled.

As in the discrete case (see Section 3.1.3), particularly useful special cases are the models, where the dynamics are independent of the parameter and given the state $\mathbf{x}(t_k)$ the prior distribution $p(\boldsymbol{\theta})$ belongs to the conjugate class of the likelihood $p(\mathbf{y}_k \,|\, \mathbf{x}(t_k),\boldsymbol{\theta})$. This is the case, for example, in estimation of spread of infectious diseases in Section 4.2 and in the noisy pendulum example in the next section.

Rao-Blackwellized Sequential Importance Resampling Smoother

Optimal smoothing of continuous-discrete conditionally Gaussian models can be performed analogously to the discrete-time case. The continuous-discrete conditionally Gaussian Rao-Blackwellized particle smoother has almost the same form as the discrete-time Rao-Blackwellized particle smoother Algorithm 3.11. The smoothed estimate for each measurement time can be obtained by considering the state sequences discretely only at measurement points and in this case the algorithm is exactly the discrete-time smoothing algorithm. But also the whole trajectories of the means, covariances and state inner processes can be stored during the simulation in Algorithm 3.31 and instead of the discrete-time Kalman smoother the continuous-discrete Kalman smoother can be used for computing the smoothed estimate for all instances of time.

The case of unknown static parameters (Storvik, 2002) is analogous to the discrete case and the smoothed distribution is the distribution obtained after all the measurements have been processed.

3.2.4 Illustrative Examples

Beneš-Daum Filtering Problem

The *Beneš-Daum filtering problem* (Daum, 1984, 1986) considered here is a *continuous-discrete* filtering problem, where the dynamic model is a scalar stochastic differential equation

$$\mathrm{d}x = \tanh(x)\,\mathrm{d}t + \mathrm{d}\beta, \tag{3.289}$$

where $\beta(t)$ is a standard Brownian motion. The measurements y_k are obtained at discrete times t_k from the model

$$y(t_k) = x(t_k) + r_k, \tag{3.290}$$

where $r_k \sim \mathrm{N}(0, \sigma^2)$. The formal solution to the filtering problem was analyzed already in Example 3.4 and several approximate filters and smoothers were given in Examples 3.7, 3.8, 3.9, 3.10 and 3.11.

The filters in the examples and couple of additional filters were tested using simulated data and the results can be summarized as follows:

- The Gaussian approximation based continuous-discrete methods, that is, the continuous-discrete unscented Kalman filter (Example 3.9) and the continuous-discrete extended Kalman filter (Example 3.7) give practically the same results as the theoretically optimal Beneš-Daum filter (Example 3.4).
- Also the Gaussian approximation based discrete-time methods, that is, the discrete-time unscented Kalman filter (Algorithm 3.7) and the discrete-time extended Kalman filter (Algorithm 3.3) give practically the same results

as the Beneš-Daum filter. In the discrete filters the dynamic model was discretized using a single step of Euler integration.

- The bootstrap filter (Algorithm 3.26) constructed for the filtering model gives almost the same result as the Gaussian approximation based UKF and EKF filters, but with moderate number of particles the result is slightly inferior to the results of the Gaussian filters.
- The continuous-discrete particle filter (Example 3.11) gives quite much the same result as the bootstrap filter with moderate and large number of particles, but with low number of particles the results are slightly inferior to the results of the bootstrap filter.

The results indicate that the non-linearity in this particular filtering problem is not very strong and for this reason Gaussian approximation based filters work very well. The dynamic model is quite linear and thus the simple Euler integration gives a good approximation. Because Gaussian approximation based filters give almost an optimal result, particle based methods can reach their performance only when quite high number of particles is used. The importance process constructed in Example 3.11 seems to be quite inefficient, because using the dynamic model as the importance process (in bootstrap filter) leads to better results with a low number of particles. One reason for the better result of the bootstrap filter can be the quite high measurement noise variance.

The conclusion is that the best choice for this kind of filtering problem would be one of the Gaussian approximation based filters.

Noisy Physical Pendulum

Consider the noisy pendulum model in Example 3.12, which has the dynamic model

$$\begin{aligned} \frac{\mathrm{d}x_1}{\mathrm{d}t} &= x_2 \\ \mathrm{d}x_2 &= -a^2 \sin(x_1)\,\mathrm{d}t + q^{1/2}\,\mathrm{d}\beta, \end{aligned} \tag{3.291}$$

Assume that the state of the pendulum is measured once per unit time and the measurements are corrupted by Gaussian measurement noise with an unknown variance σ^2. A suitable model in this case is

$$\begin{aligned} y_k &\sim \mathrm{N}(x_1(t_k), \sigma^2) \\ \sigma^2 &\sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2), \end{aligned} \tag{3.292}$$

The variance σ^2 is now an unknown static variable, where the procedure of Rao-Blackwellization can be applied:

- Assume that the posterior distribution of σ^2, given the state and measurement histories up to step $k-1$ is
$$\sigma^2 \sim \text{Inv-}\chi^2(\nu_{k-1}, \sigma^2_{k-1}). \tag{3.293}$$
where the parameters ν_{k-1} and σ^2_{k-1} are known.

- Given the state at step k, that is, $\mathbf{x}(t_k)$ the marginal distribution of the measurement y_k is Student's T:
$$\begin{aligned} p(y_k \,|\, \mathbf{x}(t_k)) &= \int \mathrm{N}(y_k \,|\, x_1(t_k), \sigma^2) \, \text{Inv-}\chi^2(\sigma^2 \,|\, \nu_{k-1}, \sigma^2_{k-1}) \, \mathrm{d}(\sigma^2) \\ &= t_{\nu_k}(y_k \,|\, x_1(t_k), \sigma^2_k), \end{aligned} \tag{3.294}$$
with parameters
$$\begin{aligned} \nu_k &= \nu_{k-1} + 1 \\ \sigma^2_k &= \frac{\nu_{k-1}\, \sigma^2_{k-1} + (y_k - x_1(t_k))^2}{\nu_{k-1} + 1}. \end{aligned} \tag{3.295}$$

- Given the measurement y_k and the state $\mathbf{x}(t_k)$, the posterior distribution of σ^2 given the state and measurement histories up to time step k is again Inv-χ^2:
$$\sigma^2 \,|\, \mathbf{x}(t_k), y_k \sim \text{Inv-}\chi^2(\nu_k, \sigma^2_k). \tag{3.296}$$

As already discussed in Example 3.12 the importance process can be formed by using either EKF or UKF such that we obtain 2-dimensional Gaussian approximation to the posterior distribution of the state $\mathbf{x}(t_k) = (x_1(t_k)\; x_2(t_k))^T$. Forming this approximation requires that the variance σ^2 is assumed to be known, but fortunately a very rough approximation based on the estimated σ^2_k is enough in practice.

The full state of the algorithm at time step $k-1$ consists of the set of particles
$$\{w^{(i)}_{k-1}, x^{(i)}_{1,k-1}, x^{(i)}_{2,k-1}, \nu^{(i)}_{k-1}, \sigma^{2,(i)}_{k-1}\} \tag{3.297}$$
where $w^{(i)}_{k-1}$ is the importance weight, $x^{(i)}_{1,k-1}, x^{(i)}_{2,k-1}$ is the state of the pendulum, and $\nu^{(i)}_{k-1}, \sigma^{2,(i)}_{k-1}$ are the sufficient statistics of the variance parameter. One step of the resulting continuous-discrete Rao-Blackwellized particle filtering algorithm is the following:

1. For each particle, perform EKF/UKF prediction from t_{k-1} to t_k and update the state by the measurement y_k. Assume that the marginal mean and covariance of $x_2(t_k)$, when the EKF/UKF was started from $x_1(t_{k-1}) = x^{(i)}_{1,k-1}, x_2(t_{k-1}) = x^{(i)}_{2,k-1}$ are $m^{(i)}_{2,k}, P^{(i)}_{22,k}$.

2. For each i simulate the scaled importance process, and the logarithm of likelihood ratio from time t_{k-1} to time t_k:

$$\begin{aligned}
\frac{\mathrm{d}s_1^{*(i)}}{\mathrm{d}t} &= s_2^{*(i)} \\
\mathrm{d}s_2^{*(i)} &= \left(\frac{m_{2,k}^{(i)} - x_{2,k-1}^{(i)}}{\sqrt{P_{22,k}^{(i)}\,\Delta t}} \right) \mathrm{d}t + q^{1/2}\,\mathrm{d}\beta \\
\mathrm{d}\lambda^{(i)} &= -\frac{a^2}{q^{1/2}} \sin(s_1^{*(i)}(t))\,\mathrm{d}\beta - \frac{(m_{2,k}^{(i)} - x_{2,k-1}^{(i)})}{\sqrt{P_{22,k}^{(i)}\,\Delta t}}\,\mathrm{d}\beta \\
&\quad - \frac{a^2}{q^{1/2}} \sin(s_1^{*(i)}(t)) \left(\frac{m_{2,k}^{(i)} - x_{2,k-1}^{(i)}}{\sqrt{P_{22,k}^{(i)}\,\Delta t}} \right) \mathrm{d}t \\
&\quad - \frac{1}{2}\frac{a^4}{q} \sin^2(s_1^{*(i)}(t))\,\mathrm{d}t - \frac{1}{2} \left(\frac{m_{2,k}^{(i)} - x_{2,k-1}^{(i)}}{\sqrt{P_{22,k}^{(i)}\,\Delta t}} \right)^2 \mathrm{d}t,
\end{aligned} \tag{3.298}$$

starting from initial conditions $s_1^{*(i)}(t_{k-1}) = x_{1,k-1}^{(i)}$, $s_2^{*(i)}(t_{k-1}) = x_{2,k-1}^{(i)}$, $\lambda^{(i)}(t_{k-1}) = 0$ and set

$$\begin{aligned}
x_{1,k}^{(i)} &= s_1^{*(i)}(t_k) \\
x_{2,k}^{(i)} &= s_2^{*(i)}(t_k) \\
Z_k^{(i)} &= \exp(\lambda^{(i)}(t_k)).
\end{aligned} \tag{3.299}$$

3. Compute the new sufficient statistics:

$$\begin{aligned}
\nu_k^{(i)} &= \nu_{k-1}^{(i)} + 1 \\
\sigma_k^{2,(i)} &= \frac{\nu_{k-1}\,\sigma_{k-1}^{2,(i)} + \left(y_k - x_{1,k}^{(i)}\right)^2}{\nu_{k-1}^{(i)} + 1}.
\end{aligned} \tag{3.300}$$

4. Compute the new weights as

$$w_k^{(i)} \propto w_{k-1}^{(i)}\, Z_k^{(i)}\, t_{\nu_k^{(i)}} \left(y_k \mid x_{1,k}^{(i)}, \sigma_k^{2,(i)} \right). \tag{3.301}$$

5. Resample if needed.

Figure 3.7 shows the result of applying the continuous-discrete particle filter with UKF proposal and 1000 particles to a simulated data. The data was generated

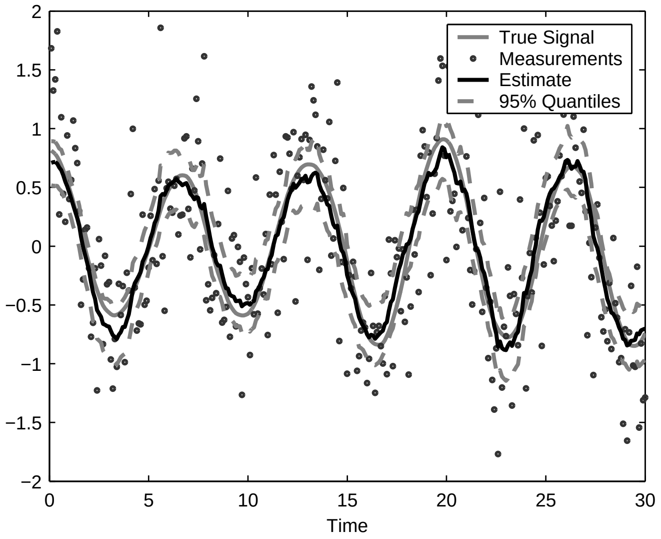

Figure 3.7: The result of applying continuous-discrete particle filter with UKF proposal to a simulated noisy pendulum data.

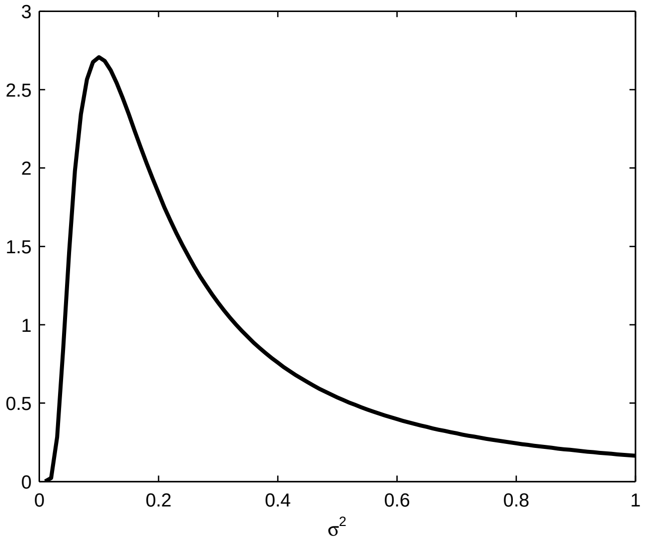

Figure 3.8: The prior distribution of variance in the noisy pendulum problem.

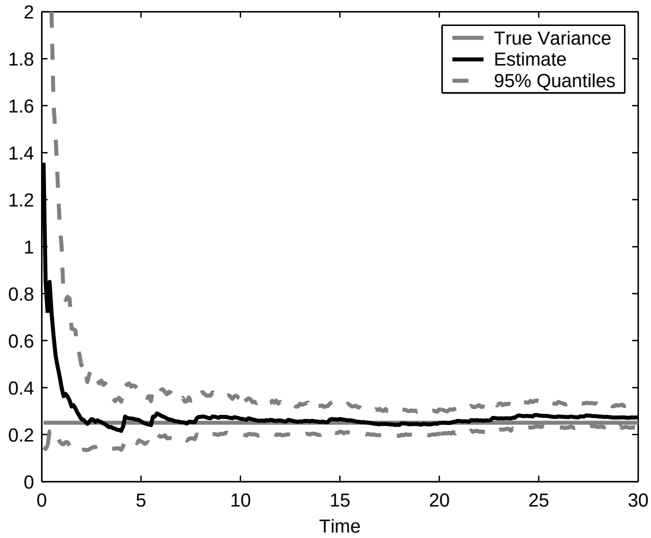

Figure 3.9: The evolution of variance distribution in the noisy pendulum problem.

from the noisy pendulum model with process noise spectral density $q = 0.01$, angular velocity $a = 1$ and the sampling step size was $\Delta t = 0.1$. The estimate can be seen to be quite close to the true signal.

In the simulation, the true measurement variance was $\sigma^2 = 0.25$. The prior distribution used for the unknown variance parameter was $\sigma^2 \sim$ Inv-$\chi^2(2, 0.2)$, which is shown in the Figure 3.8. The chosen prior distribution does support the true value, but there are also several other values which have even higher prior probability than the true value.

The evolution of the posterior distribution of the variance parameter is shown in the Figure 3.9. In the beginning the uncertainty about the variance is higher, but the distribution quickly converges to the area of the true value.

Reentry Vehicle Tracking

Here we consider the reentry tracking problem, where a radar is used for tracking a space vehicle, which enters the atmosphere at a very high speed. The purpose of including the simulation into this thesis is to test the practical applicability of the new filters proposed in this thesis into a previously proposed difficult tracking problem. The reentry problem was used for demonstrating the performance of UKF in (Julier and Uhlmann, 2004b) and slight corrections to the equations and simulation parameters were later published in (Julier and Uhlmann, 2004a).

The stochastic equations of motion for the space vehicle are given as (Julier and Uhlmann, 2004b,a)[11]

$$\begin{aligned}
R(t) &= \sqrt{x_1^2(t) + x_2^2(t)} \\
V(t) &= \sqrt{x_3^2(t) + x_4^2(t)} \\
b(t) &= b_0 \, \exp\left(x_5(t)\right) \\
D(t) &= b(t) \, \exp\left(\frac{R_0 - R}{H_0}\right) \, V(t) \\
G(t) &= -\frac{Gm_0}{R^3(t)} \\
\dot{x}_1(t) &= x_3(t) \\
\dot{x}_2(t) &= x_4(t) \\
\dot{x}_3(t) &= D(t) \, x_3(t) + G(t) \, x_1(t) + w_1(t) \\
\dot{x}_4(t) &= D(t) \, x_4(t) + G(t) \, x_2(t) + w_2(t) \\
\dot{x}_5(t) &= w_3(t),
\end{aligned} \tag{3.302}$$

where $w_1(t)$, $w_2(t)$, $w_3(t)$ are white Gaussian process noises with known joint spectral density. The constants are (Julier and Uhlmann, 2004b):

$$\begin{aligned}
b_0 &= -0.59783 \\
H_0 &= 13.406 \\
Gm_0 &= 3.9860 \times 10^5 \\
R_0 &= 6374.
\end{aligned} \tag{3.303}$$

In the article (Julier and Uhlmann, 2004b) the radar measures 10 times per second and the simulated discrete process noise covariance is

$$\mathbf{Q}(k) = \begin{bmatrix} 2.4064 \times 10^{-5} & 0 & 0 \\ 0 & 2.4064 \times 10^{-5} & 0 \\ 0 & 0 & 0 \end{bmatrix}. \tag{3.304}$$

In the actual filter implementation the lower right corner term in the modeled process noise covariance was set to the value 10^{-6} (Julier and Uhlmann, 2004a). This additional small noise term is used for the constant parameter to enhance the filter stability.

If we interpret the dynamic model (3.302) as originally continuous time model, we may assume that the discretized covariance is actually an approximation to a non-linear continuous-time process driven by continuous-time process noise with spectral density $\mathbf{Q}_c$. In this interpretation it is reasonable to assume that

[11] Here we have used the Newton's notation for derivatives $\dot{x} = \mathrm{d}x/\mathrm{d}t$.

the relationship between the discrete covariance and the continuous-time spectral density is originally the approximation $\mathbf{Q}(k) \approx \mathbf{Q}_c \, \Delta t$ where $\Delta t = 0.1\,\mathrm{s}$. Thus the true spectral density matrix is the matrix in (3.304) multiplied by 10.

The continuous-time Itô stochastic differential equation interpretation of the dynamic model (3.302) is

$$\begin{aligned}
\mathrm{d}x_1 &= x_3 \, \mathrm{d}t \\
\mathrm{d}x_2 &= x_4 \, \mathrm{d}t \\
\mathrm{d}x_3 &= D(t) \, x_3 \, \mathrm{d}t + G(t) \, x_1 \, \mathrm{d}t + \mathrm{d}\beta_1(t) \\
\mathrm{d}x_4 &= D(t) \, x_4 \, \mathrm{d}t + G(t) \, x_2 \, \mathrm{d}t + \mathrm{d}\beta_2(t) \\
\mathrm{d}x_5 &= \mathrm{d}\beta_3(t),
\end{aligned} \tag{3.305}$$

where $(\beta_1(t), \beta_2(t), \beta_3(t))$ is a Brownian motion with diffusion matrix $\mathbf{Q}_c$.

The radar is located at $(x_r, y_r) = (R_0, 0)$ and the measurement model is

$$\begin{aligned}
r_k &= \sqrt{(x_1(t_k) - x_r)^2 + (x_2(t_k) - y_r)^2} + e_k^r \\
\theta_k &= \tan^{-1}\left(\frac{x_2(t_k) - y_r}{x_1(t_k) - x_r}\right) + e_k^a,
\end{aligned} \tag{3.306}$$

where the $e_k^r \sim \mathrm{N}(0, \sigma_r^2)$ and $e_k^a \sim \mathrm{N}(0, \sigma_a^2)$.

The initial distribution of the state is assumed to be multidimensional Gaussian with mean and covariance (Julier and Uhlmann, 2004b)

$$\begin{aligned}
\mathbf{m}_0 &= \begin{pmatrix} 6500.4 \\ 349.14 \\ -1.8093 \\ -6.7967 \\ 0 \end{pmatrix} \\
\mathbf{P}_0 &= \begin{pmatrix} 10^{-6} & 0 & 0 & 0 & 0 \\ 0 & 10^{-6} & 0 & 0 & 0 \\ 0 & 0 & 10^{-6} & 0 & 0 \\ 0 & 0 & 0 & 10^{-6} & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.
\end{aligned} \tag{3.307}$$

In simulations, the initial state was drawn from Gaussian distribution with the following mean and covariance:

$$
\begin{aligned}
\mathbf{m}_0^{\text{sim}} &= \begin{pmatrix} 6500.4 \\ 349.14 \\ -1.8093 \\ -6.7967 \\ 0.6932 \end{pmatrix} \\
\mathbf{P}_0^{\text{sim}} &= \begin{pmatrix} 10^{-6} & 0 & 0 & 0 & 0 \\ 0 & 10^{-6} & 0 & 0 & 0 \\ 0 & 0 & 10^{-6} & 0 & 0 \\ 0 & 0 & 0 & 10^{-6} & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.
\end{aligned}
\tag{3.308}
$$

In order to test the *performance of different types of filters*, continuous-discrete extended Kalman filter, continuous-discrete unscented Kalman filter, continuous-discrete bootstrap filter and continuous-discrete sequential importance resampling filter were implemented and tested with simulated data. The importance process for the CD-SIR was constructed in analogous manner to the Example 3.11. The same standard deviations as in (Chang et al., 1977) (which Julier and Uhlmann (2004b) also refers to) were used in the simulation:

$$
\begin{aligned}
\sigma_r &= 10^{-3}\,\mathrm{km} \\
\sigma_a &= 0.17\,\mathrm{mrad},
\end{aligned}
\tag{3.309}
$$

and the state was measured once per second. The results can be summarized as follows:

- The EKF gives the best result, which is most likely due to the very low process and measurement noises. That is, if one filter should be selected for a real implementation, it would be EKF.
- The UKF gives quite much the same result as EKF, but for some reason it gives a slightly higher error than EKF. However, the implementation of UKF in this case is much easier than of EKF, because the Jacobian matrices of the drift and measurement functions are quite complicated expressions.
- The bootstrap filter gives the worst result and it diverges in many of the test cases. The reason for this might be that the measurements are very accurate and thus many of the predicted particles “miss” the next measurement on each step, which leads to divergence with significantly non-zero probability.
- The continuous-discrete importance resampling filter gives a result, which is worse than the result of EKF and UKF, but which is quite close to the truth. As opposed to the bootstrap filter the continuous-discrete importance

resampling filter does not diverge, because the importance process keeps it approximately on the track. However, the importance process is not a very good one, because the results are still worse than of EKF and UKF.

In order to test the *effect of discretization* we increased the process noise by factor of 100 (factor 10 in standard error scale) and the process noise was

$$\mathbf{Q}_c = \begin{bmatrix} 0.0241 & 0 & 0 \\ 0 & 0.0241 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \tag{3.310}$$

The above diffusion matrix was used in simulation, but in the filter we used slightly different diffusion matrix $\mathbf{Q}'_c$, following the convention in (Julier and Uhlmann, 2004a)

$$\mathbf{Q}'_c = \begin{bmatrix} 0.0241 & 0 & 0 \\ 0 & 0.0241 & 0 \\ 0 & 0 & 10^{-5} \end{bmatrix}. \tag{3.311}$$

The simulated data were generated by simulating the stochastic differential equation (3.305) with 100 steps of Euler-Maruyama scheme (Kloeden and Platen, 1999) between each measurement.

The standard deviations of the measurements in simulation were selected to be

$$\begin{aligned} \sigma_r &= 0.01\,\mathrm{km} \\ \sigma_a &= 1.7\,\mathrm{mrad}, \end{aligned} \tag{3.312}$$

which are 10 times the standard deviations in (Chang et al., 1977). The standard deviations were selected to be higher in order to make the problem harder and to reveal the differences between the different approximations.

The discrete UKF (DUKF) and continuous-discrete UKF (CDUKF) were both implemented using 10 steps of Runge-Kutta integration between measurements, but the difference was in handling of process noise:

- In DUKF, each measurement is processed as follows:

    1. Integrate each of the sigma point through the noise free dynamic model using 10 steps of the Runge-Kutta integration.
    2. Compute the predicted and mean covariance, and model process noise effect by approximating the discrete covariance by $\mathbf{Q}'(k) \approx \mathbf{Q}'_c\,\Delta t$.
    3. Perform standard UKF update step for the measurement.

- In CDUKF, each measurement is processed as follows:

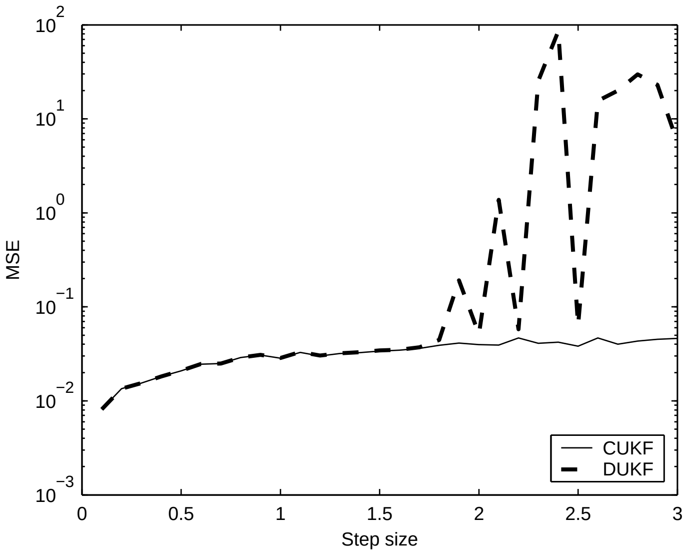

Figure 3.10: Mean squared error (MSE) versus time step size in the reentry filtering problem. Results are from 200 Monte Carlo runs per time step with the continuous-discrete UKF (CDUKF) and discrete UKF (DUKF). The higher errors of DUKF with longer time step sizes are caused by numerical problems on the prediction step.

1. Integrate the mean and covariance differential equations using 10 steps of the Runge-Kutta integration, and using $\mathbf{Q}'_c$ as the diffusion matrix (or spectral density) of the process noise.
2. Perform standard UKF update step for the measurement.

The amount of computations required by the CDUKF is slightly higher than of DUKF, but the number of evaluations of the dynamic model function is the same for both the models and thus the practical total difference is small. The algorithm parameters in the unscented transforms were selected to be $\alpha = 1/2, \beta = 2, \kappa = -2$.

Simulations were performed using different time steps of $\Delta t = 0.1, 0.2, \ldots, 3$ seconds and the results from 200 Monte Carlo simulations per step size are shown in Figure 3.10. The continuous-discrete UKF (CDUKF) and discrete UKF (DUKF) have very much the same performance when the time step size is short. However, when the time step grows, DUKF encounters numerical problems and its error grows rapidly. At the same time, no numerical problems can be seen in the CDUKF and its error grows much slower.

In this simulation scenario the advantage of the continuous-time formulation over the discrete-time formulation is the numerical stability. This stability is due to that when the noise process is modeled as a continuous-time process the non-linear dynamic model cannot force the covariance to become non-positive definite. In the discrete-time formulation the covariance can become non-positive definite.

In this particular problem it is not essential whether uncertainties are modeled as discrete-time or continuous-time stochastic processes. Since the performance of CDUKF seems to be at least that of DUKF, in cases where the continuous-time stochastic process formulation is more accurate in modeling point of view, the continuous-discrete UKF is likely to perform better than the discrete UKF.

3.3 Continuous-Time Filtering and Smoothing

This section presents the theory of Gaussian approximation based non-linear optimal filtering and smoothing. The novel contribution of this section is the continuous-time unscented Kalman-Bucy filter.

3.3.1 Continuous-Time Filtering Model

The most general form of the continuous-time filtering models considered in this thesis is

$$\begin{aligned} \mathrm{d}\mathbf{x}(t) &= \mathbf{f}(\mathbf{x}(t), t)\,\mathrm{d}t + \mathbf{L}(t)\,\mathrm{d}\boldsymbol{\beta}(t) \\ \mathrm{d}\mathbf{y}(t) &= \mathbf{h}(\mathbf{x}(t), t)\,\mathrm{d}t + \mathbf{V}(t)\,\mathrm{d}\boldsymbol{\eta}(t), \end{aligned} \tag{3.313}$$

where

- $\mathbf{x}(t) \in \mathbb{R}^n$ is the state process,
- $\mathbf{y}(t) \in \mathbb{R}^m$ is the (integrated) measurement process,
- $\mathbf{f}$ is the drift function,
- $\mathbf{h}$ is the measurement model function,
- $\mathbf{L}(t)$ and $\mathbf{V}(t)$ are arbitrary time varying matrices, independent of $\mathbf{x}(t)$ and $\mathbf{y}(t)$,
- $\boldsymbol{\beta}(t)$ and $\boldsymbol{\eta}(t)$ are independent Brownian motions with diagonal diffusion matrices $\mathbf{Q}_c(t)$ and $\mathbf{R}_c(t)$, respectively.

The filtering model can also be formulated in terms of formal white noises $\mathbf{w}(t) = \mathrm{d}\boldsymbol{\beta}(t)/\mathrm{d}t$, $\mathbf{e}(t) = \mathrm{d}\boldsymbol{\eta}(t)/\mathrm{d}t$, and differential measurement $\mathbf{z}(t) = \mathrm{d}\mathbf{y}(t)/\mathrm{d}t$ as follows (Jazwinski, 1970):

$$\begin{aligned} \frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} &= \mathbf{f}(\mathbf{x}(t), t) + \mathbf{L}(t)\,\mathbf{w}(t) \\ \mathbf{z}(t) &= \mathbf{h}(\mathbf{x}(t), t) + \mathbf{V}(t)\,\mathbf{e}(t), \end{aligned} \tag{3.314}$$

where the white noise processes $\mathbf{w}(t)$ and $\mathbf{e}(t)$ have spectral densities $\mathbf{Q}_c(t)$ and $\mathbf{R}_c(t)$, respectively.

3.3.2 Kalman-Bucy Filtering and Smoothing

The Kalman-Bucy filter (Kalman and Bucy, 1961) is the formal solution to the linear Gaussian optimal filtering problem

$$\begin{aligned} \mathrm{d}\mathbf{x} &= \mathbf{F}(t)\,\mathbf{x}\,\mathrm{d}t + \mathbf{L}(t)\,\mathrm{d}\boldsymbol{\beta} \\ \mathrm{d}\mathbf{y} &= \mathbf{H}(t)\,\mathbf{x}\,\mathrm{d}t + \mathbf{V}(t)\,\mathrm{d}\boldsymbol{\eta}, \end{aligned} \tag{3.315}$$

where

- $\mathbf{x}(t) \in \mathbb{R}^n$ is the state process,
- $\mathbf{y}(t) \in \mathbb{R}^m$ is the measurement process,
- $\mathbf{F}(t)$ is the dynamic model matrix,
- $\mathbf{H}(t)$ is the measurement model matrix,
- $\mathbf{L}(t)$ and $\mathbf{V}(t)$ are arbitrary time varying matrices, independent of $\mathbf{x}(t)$ and $\mathbf{y}(t)$,
- $\boldsymbol{\beta}(t)$ and $\boldsymbol{\eta}(t)$ are independent Brownian motions with diagonal diffusion matrices $\mathbf{Q}_c(t)$ and $\mathbf{R}_c(t)$, respectively.

The solution is given as follows:

Theorem 3.8 (Kalman-Bucy filter)**.** *The optimal filter, which computes the posterior distribution $p(\mathbf{x}(t) \mid \mathscr{Y}_t) = \mathrm{N}(\mathbf{x}(t) \mid \mathbf{m}(t), \mathbf{P}(t))$ for the system* (3.315) *is given as*

$$\begin{aligned}
\mathbf{K}_c(t) &= \mathbf{P}(t)\,\mathbf{H}^T(t)\,\left[\mathbf{V}(t)\,\mathbf{R}_c(t)\,\mathbf{V}^T(t)\right]^{-1} \\
\frac{\mathrm{d}\mathbf{m}(t)}{\mathrm{d}t} &= \mathbf{F}(t)\,\mathbf{m}(t) + \mathbf{K}_c(t)\,\left[\mathbf{z}(t) - \mathbf{H}(t)\,\mathbf{m}(t)\right] \\
\frac{\mathrm{d}\mathbf{P}(t)}{\mathrm{d}t} &= \mathbf{F}(t)\,\mathbf{P}(t) + \mathbf{P}(t)\,\mathbf{F}^T(t) + \mathbf{L}(t)\,\mathbf{Q}_c(t)\,\mathbf{L}^T(t) \\
&\quad - \mathbf{K}_c(t)\,\mathbf{V}(t)\,\mathbf{R}_c(t)\,\mathbf{V}^T(t)\,\mathbf{K}_c^T(t),
\end{aligned} \tag{3.316}$$

where $\mathbf{z}(t) = \mathrm{d}\mathbf{y}(t)/\mathrm{d}t$ is the formal *derivative of the measurement process.*

Proof. See Appendix A.3.1. □

The equations of the *continuous-time Kalman-Bucy smoother* are exactly the same as the equations of the continuous-discrete Kalman smoother. That is, the smoother equations are the ones given in the Algorithm 3.18, except that $\mathbf{m}(t)$ and $\mathbf{P}(t)$ are defined to be the mean and covariance computed by the Kalman-Bucy filter.

3.3.3 Continuous-Time Gaussian Approximations

Extended Kalman-Bucy Filter and Smoother

The extended Kalman-Bucy filter (see, e.g., Gelb, 1974) is an extension of the Kalman-Bucy filter to non-linear models of the form (3.313).

Algorithm 3.32 (Extended Kalman-Bucy filter)**.** *The equations of the Extended Kalman-Bucy filter (EKBF) are:*

$$\begin{aligned}\mathbf{K}_c(t) &= \mathbf{P}(t)\,\mathbf{H}^T(\mathbf{m}(t),t)\;\left[\mathbf{V}(t)\,\mathbf{R}_c(t)\,\mathbf{V}^T(t)\right]^{-1} \\ \frac{\mathrm{d}\mathbf{m}(t)}{\mathrm{d}t} &= \mathbf{f}(\mathbf{m}(t),t) + \mathbf{K}_c(t)\;[\mathbf{z}(t) - \mathbf{H}(\mathbf{m}(t),t)\,\mathbf{m}(t)] \\ \frac{\mathrm{d}\mathbf{P}(t)}{\mathrm{d}t} &= \mathbf{F}(\mathbf{m}(t),t)\,\mathbf{P}(t) + \mathbf{P}(t)\,\mathbf{F}^T(\mathbf{m}(t),t) + \mathbf{L}(t)\,\mathbf{Q}_c(t)\,\mathbf{L}^T(t) \\ &\quad - \mathbf{K}_c(t)\,\mathbf{V}(t)\,\mathbf{R}_c(t)\,\mathbf{V}^T(t)\,\mathbf{K}_c^T(t),\end{aligned} \tag{3.317}$$

where $\mathbf{z}(t) = \mathrm{d}\mathbf{y}(t)/\mathrm{d}t$ *is the* formal *derivative of the measurement process* $\mathbf{y}(t)$, $\mathbf{F}(\cdot)$ *is the Jacobian matrix of* $\mathbf{f}(\cdot)$ *with elements* $F_{ij} = \partial f_i/\partial x_j$*, and* $\mathbf{H}(\cdot)$ *is the Jacobian matrix of* $\mathbf{h}(\cdot)$ *with elements* $H_{ij} = \partial h_i/\partial x_j$.

This filter is actually the first order EKBF, but analogously to the discrete case and continuous-discrete case, second order approximations could be applied as well.

The equations of the *continuous-time extended Kalman-Bucy smoother* are exactly the same as the equations of the continuous-discrete extended Kalman smoother in Algorithm 3.23, but with the difference that $\mathbf{m}(t)$ and $\mathbf{P}(t)$ are the mean and covariance computed by the (continuous-time) extended Kalman-Bucy filter.

Unscented Kalman-Bucy Filter and Smoother

By taking the formal limit of the discrete-time unscented Kalman filter equations in Algorithm 3.7, the following novel continuous-time filter can be derived:

Theorem 3.9 (Unscented Kalman-Bucy filter)**.** *The stochastic differential equations corresponding to the UKF in the continuous-time limit of state and measurement processes, that is, the unscented Kalman-Bucy filter equations, are given as*

$$\mathbf{X}(t) = \begin{bmatrix}\mathbf{m}(t) & \cdots & \mathbf{m}(t)\end{bmatrix} + \sqrt{c}\,\begin{bmatrix}\mathbf{0} & \sqrt{\mathbf{P}(t)} & -\sqrt{\mathbf{P}(t)}\end{bmatrix} \tag{3.318}$$

$$\mathbf{K}_c(t) = \mathbf{X}(t)\,\mathbf{W}\,\mathbf{h}^T(\mathbf{X}(t),t)\left[\mathbf{V}(t)\,\mathbf{R}_c(t)\,\mathbf{V}^T(t)\right]^{-1} \tag{3.319}$$

$$\begin{aligned}\frac{\mathrm{d}\mathbf{P}(t)}{\mathrm{d}t} &= \mathbf{X}(t)\,\mathbf{W}\,\mathbf{f}^T(\mathbf{X}(t),t) + \mathbf{f}(\mathbf{X}(t),t)\,\mathbf{W}\,\mathbf{X}^T(t) \\ &\quad + \mathbf{L}(t)\,\mathbf{Q}_c(t)\,\mathbf{L}^T(t) \\ &\quad - \mathbf{K}_c(t)\,\mathbf{V}(t)\,\mathbf{R}_c(t)\,\mathbf{V}^T(t)\,\mathbf{K}_c^T(t)\end{aligned} \tag{3.320}$$

$$\frac{\mathrm{d}\mathbf{m}(t)}{\mathrm{d}t} = \mathbf{f}(\mathbf{X}(t),t)\,\mathbf{w}_m + \mathbf{K}_c(t)\;[\mathbf{z}(t) - \mathbf{h}(\mathbf{X}(t),t)\,\mathbf{w}_m]\,, \tag{3.321}$$

where we have formally defined the differential measurement $\mathbf{z}(t) = \mathrm{d}\mathbf{y}(t)/\mathrm{d}t$. *In terms of the process* $\mathbf{y}(t)$ *the mean equation can be written as Itô stochastic* *differential equation*

$$\mathrm{d}\mathbf{m}(t) = [\mathbf{f}(\mathbf{X}(t), t) - \mathbf{K}_c(t)\,\mathbf{h}(\mathbf{X}(t), t)]\;\mathbf{w}_m\,\mathrm{d}t + \mathbf{K}_c(t)\,\mathrm{d}\mathbf{y}(t). \tag{3.322}$$

Proof. See Appendix A.3.3. □

Corollary 3.1 (Prediction differential equations)**.** *The approximate predicted mean* $\mathbf{m}(t)$ *and covariance* $\mathbf{P}(t)$ *of the state for times* $t \geq t_0$ *given the mean and covariance at the time instance* t_0 *can be computed by integrating the differential equations*

$$\begin{aligned}
\mathbf{X}(t) &= \begin{bmatrix}\mathbf{m}(t) & \cdots & \mathbf{m}(t)\end{bmatrix} + \sqrt{c}\,\begin{bmatrix}\mathbf{0} & \sqrt{\mathbf{P}(t)} & -\sqrt{\mathbf{P}(t)}\end{bmatrix} && (3.323)\\
\frac{\mathrm{d}\mathbf{P}(t)}{\mathrm{d}t} &= \mathbf{X}(t)\,\mathbf{W}\,\mathbf{f}^T(\mathbf{X}(t), t) + \mathbf{f}(\mathbf{X}(t), t)\,\mathbf{W}\,\mathbf{X}^T(t)\\
&\quad + \mathbf{L}(t)\,\mathbf{Q}_c(t)\,\mathbf{L}^T(t) && (3.324)\\
\frac{\mathrm{d}\mathbf{m}(t)}{\mathrm{d}t} &= \mathbf{f}(\mathbf{X}(t), t)\,\mathbf{w}_m, && (3.325)
\end{aligned}$$

from initial conditions $\mathbf{m}(t_0)$ *and* $\mathbf{P}(t_0)$ *to time instance* t*.*

Proof. Formally set $\mathbf{R}_c(t) = \infty\,\mathbf{I}$ in the Theorem 3.9, which results in $\mathbf{K}_c(t) = \mathbf{0}$. □

The equations of the *continuous-time unscented Kalman-Bucy smoother* are exactly the same as the equations of the continuous-discrete unscented Kalman smoother in Algorithm 3.25.

3.3.4 Illustrative Examples

Beneš Filtering Problem

In the *Beneš filtering problem* (Beneš, 1981) considered here a scalar signal process $x(t)$ is observed through a scalar measurement process $y(t)$ and the models for these processes are given as

$$\begin{aligned}
\mathrm{d}x(t) &= \tanh(x(t))\,\mathrm{d}t + \mathrm{d}\beta(t) && (3.326)\\
\mathrm{d}y(t) &= \mathrm{d}x(t) + \mathrm{d}\eta(t), && (3.327)
\end{aligned}$$

where $\beta(t)$ and $\eta(t)$ are standard Brownian motions, and $x(0) = 0, y(0) = 0$.

The following estimation methods are tested:

- *Beneš filter (BF).* The exact equations for the sufficient statistics $\sigma(t), \mu(t)$ of the posterior distribution are (Beneš, 1981):

$$\begin{aligned}
\mathrm{d}\mu(t) &= \sigma(t)\,\mathrm{d}y(t) - \sigma(t)\,\mu(t)\,\mathrm{d}t\\
\mathrm{d}\sigma(t)/\mathrm{d}t &= 1 - \sigma^2(t),
\end{aligned} \tag{3.328}$$

with $\mu(0) = 0, \sigma(0) = 0$. The posterior distribution is then of the form

$$p(x(t) \,|\, \mathscr{Y}_t) \propto \cosh(x(t)) \, \exp\left\{ -\frac{(x(t) - \mu(t))^2}{2\sigma(t)} \right\}, \tag{3.329}$$

and the posterior mean can be computed as

$$\mathrm{E}[x(t) \,|\, \mathscr{Y}_t] = \mu(t) + \tanh(\mu(t)) \, \sigma(t). \tag{3.330}$$

- *Unscented Kalman-Bucy filter (UKBF).* The filtering equations are given as

$$\begin{aligned} \mathbf{X}(t) &= \begin{bmatrix} m(t) & m(t) + \sqrt{c\,P(t)} & m(t) - \sqrt{c\,P(t)} \end{bmatrix} \\ \mathbf{F}(t) &= \tanh(\mathbf{X}(t)) \\ \mathrm{d}m(t) &= \mathbf{F}(t)\, \mathbf{w}_m \, \mathrm{d}t + P(t)\,(\mathrm{d}y(t) - m(t)\,\mathrm{d}t) \\ \frac{\mathrm{d}P(t)}{\mathrm{d}t} &= 2\,\mathbf{F}(t)\,\mathbf{W}\,\mathbf{X}^T(t) + 1 - P^2(t). \end{aligned} \tag{3.331}$$

- *Extended Kalman-Bucy filter (EKBF).* The filtering equations are

$$\begin{aligned} \mathrm{d}m(t) &= \tanh(m(t))\,\mathrm{d}t + P(t)\,(\mathrm{d}y(t) - m(t)\,\mathrm{d}t) \\ \frac{\mathrm{d}P(t)}{\mathrm{d}t} &= 2\,(1 - \tanh^2(m(t)))\,P(t) + 1 - P^2(t). \end{aligned} \tag{3.332}$$

- *Linearized Kalman-Bucy filter (KBF).* As a base line solution the following linearized Kalman-Bucy filter is used. It differs from EKBF in the mean equation:

$$\begin{aligned} \mathrm{d}m(t) &= \left(1 - \tanh^2(m(t))\right)\, m(t)\,\mathrm{d}t \\ &\qquad + P(t)\,(\mathrm{d}y(t) - m(t)\,\mathrm{d}t) \\ \frac{\mathrm{d}P(t)}{\mathrm{d}t} &= 2\,(1 - \tanh^2(m(t)))\,P(t) + 1 - P^2(t). \end{aligned} \tag{3.333}$$

- *Unscented Kalman-Bucy smoother (UKBS).* The smoothing equations are given as

$$\begin{aligned} \mathbf{F}(t) &= \tanh(\mathbf{X}(t)) \\ \frac{\mathrm{d}m^s(t)}{\mathrm{d}t} &= \mathbf{F}(t)\,\mathbf{w}_m + \left(\frac{\mathbf{F}(t)\,\mathbf{W}\,\mathbf{X}^T(t) + 1}{\mathbf{X}(t)\,\mathbf{W}\,\mathbf{X}^T(t)} \right) \\ &\qquad \times (m^s(t) - \mathbf{X}(t)\,\mathbf{w}_m) \\ \frac{\mathrm{d}P^s(t)}{\mathrm{d}t} &= 2 \left(\frac{\mathbf{F}(t)\,\mathbf{W}\,\mathbf{X}^T(t) + 1}{\mathbf{X}(t)\,\mathbf{W}\,\mathbf{X}^T(t)} \right) P^s(t) - 1, \end{aligned} \tag{3.334}$$

where $\mathbf{X}(t)$ are the sigma points from the UKBF.

- *Extended Kalman-Bucy smoother (EKBS).* The smoothing equations are

$$\begin{aligned}\frac{\mathrm{d}m^s(t)}{\mathrm{d}t} &= \tanh(m(t)) + \left(\frac{(1-\tanh^2(m(t)))+1}{P(t)}\right) \\ &\quad \times (m^s(t)-m(t)) \\ \frac{\mathrm{d}P^s(t)}{\mathrm{d}t} &= 2\left(\frac{(1-\tanh^2(m(t)))+1}{P(t)}\right) P^s(t) - 1,\end{aligned} \tag{3.335}$$

  where $m(t)$ and $P(t)$ are the mean and covariance from the EKBF.

- *Linearized Kalman-Bucy smoother (KBS).* The equations for this base line smoother are

$$\begin{aligned}\frac{\mathrm{d}m^s(t)}{\mathrm{d}t} &= \left(1-\tanh^2(m(t))\right)\, m^s(t) \\ &\quad + (1/P(t))\,(m^s(t)-m(t)) \\ \frac{\mathrm{d}P^s(t)}{\mathrm{d}t} &= 2\left(\frac{(1-\tanh^2(m(t)))+1}{P(t)}\right) P^s(t) - 1,\end{aligned} \tag{3.336}$$

  where $m(t)$ and $P(t)$ are the mean and covariance from the linearized Kalman-Bucy filter.

The results from 10000 simulated data realizations are shown in Table 3.2. The values are the means of the root mean squared error (RMSE) values averaged over the realizations and the standard deviations of the RMSE means (RMSE-SE). It can be seen that the results of Beneš, EKBF and UKBF filters are practically the same and the differences are not significant when the standard errors are taken into account. The KBF gives slightly worse results, most likely due to its more inaccurate mean propagation approximation.

As could be expected, the results of smoothers are all better than of the corresponding filters and again KBS gives worse results than EKBS and UKBS. Still it can be concluded that the performances of UKBF and UKBS are at least as good as performances of EKBF and EKBS.

Continuous-Time Adaptive Control

In this section we shall test the performance of the unscented Kalman-Bucy filter in a continuous-time recursive system identification problem, which is much similar to the identification problems that arise in context of adaptive control (Åström and Wittenmark, 1995). The system has input-output transfer function

$$H(s) = \frac{1}{s^2 + a\,s} \tag{3.337}$$

Table 3.2: RMSE values of signal estimates in the Beneš filtering/smoothing problem averaged over 10000 Monte Carlo runs and standard errors of the estimated RMSE values.

Model	*RMSE*	*RMSE-SE*
BF	32.13	0.07
UKBF	32.22	0.07
EKBF	32.27	0.07
KBF	39.23	0.09
UKBS	21.92	0.05
EKBS	21.98	0.05
KBS	23.51	0.05

where the parameter a is unknown. The purpose is to recursively estimate the unknown parameter a using noisy measurements when the system is probed with a known input signal $u(t)$.

The input to the system consists of an unknown disturbance, which is modeled as a white noise process $w(t)$ with spectral density $q = 1$, and a known input signal $u(t)$ which is chosen to be a square wave with values $\{-1, 1\}$ and period of 2 time units. The Laplace transform $X(s)$ of the signal process $x(t)$ is

$$X(s) = H(s)\,W(s) + H(s)\,U(s), \tag{3.338}$$

where $W(s)$ is the Laplace transform the white noise process and $U(s)$ is the Laplace transform of the known input signal. For estimation purposes the system model is written as state space model

$$\begin{aligned} \mathrm{d}x_1(t)/\mathrm{d}t &= x_2(t) \\ \mathrm{d}x_2(t)/\mathrm{d}t &= -a(t)\,x_2(t) + w(t) + u(t) \\ \mathrm{d}a/\mathrm{d}t &= w_a(t), \end{aligned} \tag{3.339}$$

where $w_a(t)$ is a white noise process with small spectral density $q_a = 10^{-6}$, which is included to enhance the filter stability. Both the signal $x_1(t)$ and its derivative $x_2(t) = \mathrm{d}x_1(t)/\mathrm{d}t$ can be measured, but the measurements are corrupted by white noise processes $e_1(t)$ and $e_2(t)$ with spectral densities $r_1 = 0.1$ and $r_2 = 0.1$, respectively:

$$\begin{aligned} z_1(t) &= x_1(t) + e_1(t) \\ z_2(t) &= x_2(t) + e_2(t). \end{aligned} \tag{3.340}$$

In simulation, the initial conditions for both the signal x_1 and derivative were drawn randomly from normal distribution with zero mean and unit variance. The

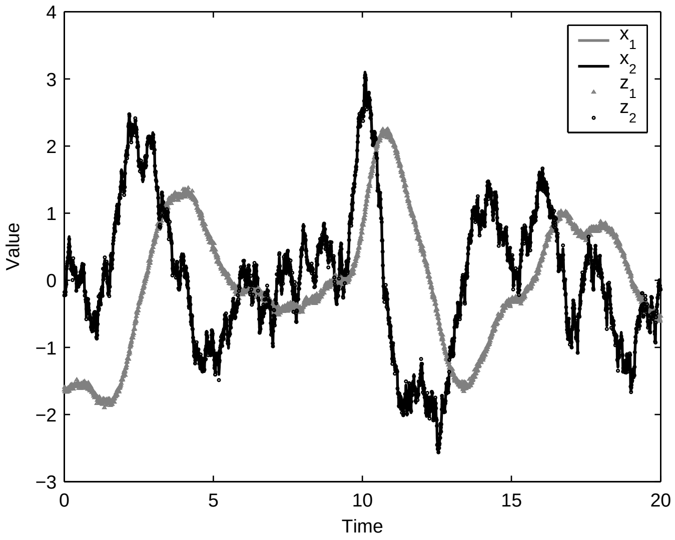

Figure 3.11: Single realization of the signal process and measurements in the parameter adaptation problem.

true value of parameter in the simulation was $a = 0.2$. The simulation was performed over time period $t \in [0, 20]$ with Euler integration and with time steps of $\Delta t = 0.01$. Single realization of the process is shown in the Figure 3.11.

The following estimators were tested:

- EKBF: First order extended Kalman-Bucy filter, where a first order Taylor series approximation is used as approximation to the state space model (3.339).
- EKBF2: Second order extended Kalman-Bucy filter, where the second order terms in the state space model (3.339) are also taken into account.
- UKBF: Unscented Kalman-Bucy filter.

Single estimation result using each of the filters is shown in Figure 3.12. It is interesting to notice that the results of EKBF2 and UKBF are actually identical. It seems that both the filters are able to take into account the effect of the second order terms exactly and because of this the estimation results are identical. The estimation result of first order EKBF seems to converge a bit more slowly, but also its estimate seems to approach the correct value. Table 3.3 shows RMSE values of the estimates of parameter a and signal $x_1(t)$ averaged over 100 Monte Carlo simulations.

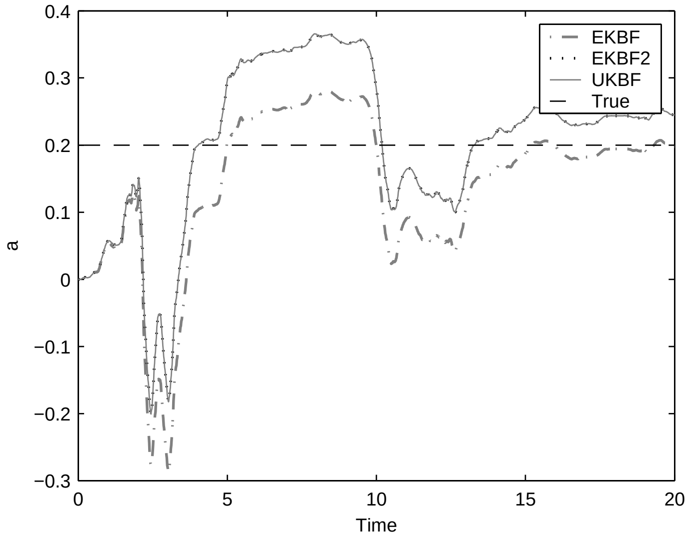

Figure 3.12: Single parameter estimation result using each of the filters.

Table 3.3: RMSE values of the parameter and signal averaged over 100 Monte Carlo runs in the parameter adaptation problem.

Model	*RMSE[a]*	*RMSE[$x_1(t)$]*
EKBF	0.23	0.11
EKBF2	0.22	0.11
UKBF	0.22	0.11

Chapter 4

Case Studies

This chapter presents applications of continuous-discrete filtering. Although many of the data in this chapter are simulated, the difference to the examples presented in the previous chapter is that the applications in this chapter are more problem oriented than in the previous chapter. This means that the purpose here is to solve existing problems, not only demonstrate the filtering methods. The following applications are presented:

- Rao-Blackwellized particle filtering based solution to *multiple target tracking in the case of an unknown number of targets* is presented in Section 4.1. Most of the contents of the section have previously been published in (Särkkä et al., 2004a, 2006b).

- Rao-Blackwellized particle filtering based solution to the *estimation of the spread of an infectious disease* is presented in Section 4.2.

- Kalman filtering and smoothing based solution to the *CATS time series prediction competition* (Lendasse et al., 2004) is presented in Section 4.3. Most parts of the section have previously been published in (Särkkä et al., 2004b, 2006a).

4.1 Multiple Target Tracking

This section presents a Rao-Blackwellized particle filter based solution to the problem of tracking an unknown number of targets using multiple sensors. The simpler case of a known number of targets was presented in the article (Särkkä et al., 2004a) and the extension to an unknown number of targets was presented in (Särkkä et al., 2006b).

4.1.1 Overview of the Problem

The basic tracking scenario consists of *sensors*, which produce noisy measurements, for example, azimuth angle measurements as illustrated in Figure 4.1. The purpose of tracking algorithm is to determine the *target trajectory* using the sensor measurements. There is additional *prior information* on the dynamics of targets, which restricts the forms of target trajectories into those that are possible when the laws of physics are taken into account.

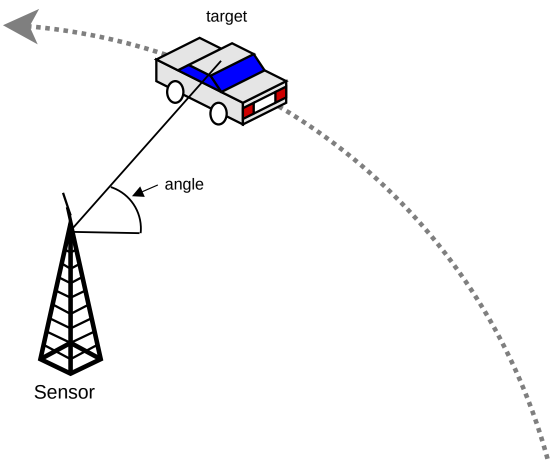

Figure 4.1: *Sensor generates angle measurements of the target, and the purpose is to determine the target trajectory.*

More general tracking scenario consists of *multiple sensors*, which may have different precisions, and they can produce different kinds of measurements, also other than azimuth angle measurements. Typically, sensors are not synchronized and they produce measurements during irregular intervals. Figure 4.2 illustrates the case of multiple sensors. In estimation point of view, increasing number of sensors will ease the estimation procedure, since we get more information on the same target trajectory.

In case of multiple targets there is an additional difficulty, because without additional information we do not know which of the measurements correspond to which targets. Figure 4.3 illustrates this problem – if the observed information are the angle measurements 1 – 4, how do we know which targets they belong to? This is called the problem of *data association*. The same problem applies to *false alarm or clutter measurements*, since we do not know if a given measurement was false alarm or a measurement from one of the targets.

Also the *number of targets* is unknown in all realistic multiple target tracking

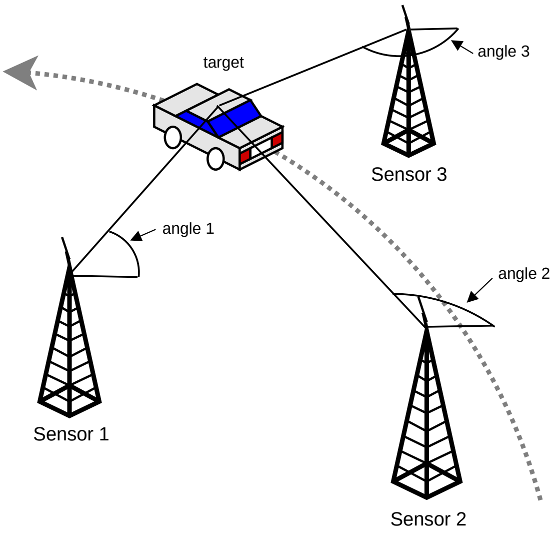

Figure 4.2: *Multiple sensors give us more information on the same trajectory and thus ease the estimation procedure.*

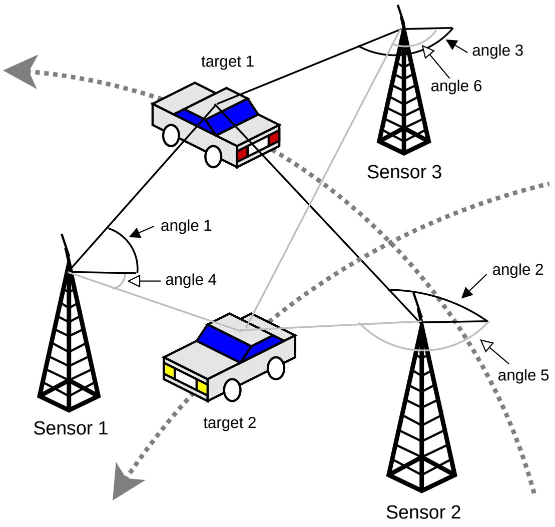

Figure 4.3: *In case of multiple targets, it is impossible to know without any additional information, which target produced which measurement.*

scenarios and it has to be estimated also.

Approaches to Data Association

The classical data association methods for multiple target tracking can be divided into two main classes (Blackman and Popoli, 1999). *Unique-neighbor data association* methods, such as *multiple hypothesis tracking* (MHT), associate each measurement with one of the previously established tracks. *All-neighbors data association* methods, such as *joint probabilistic data association* (JPDA), use all measurements for updating all the track estimates.

The idea of MHT (Bar-Shalom and Li, 1995; Blackman and Popoli, 1999; Stone et al., 1999) is to associate each measurement with one of the existing tracks, or to form a new track from the measurement. Because this association is not necessarily unique, several hypotheses are continuously formed and maintained. The MHT algorithm calculates the likelihoods of the measurements and the posterior probabilities of the hypotheses, storing only the most probable hypotheses. To enhance the computational efficiency, heuristic methods such as gating, hypothesis merging, clustering and several other strategies can be employed.

The *probabilistic multiple hypothesis tracking* (PMHT) (Streit and Luginbuhl, 1994) is a modification of the MHT, where the data associations are assumed to be independent over the target tracks. This way the computational complexity of the method is substantially reduced, but it is also impossible to model certain practical constraints, for example, to restrict the number of contacts per target to one on each scan. The RBMCDA method (Särkkä et al., 2004a) in its original form had this same restriction, but it can be easily overcome by allowing dependencies in the data association priors, as shown later in this section.

JPDA (Bar-Shalom and Li, 1995; Blackman and Popoli, 1999) approximates the posterior distributions of the targets as separate Gaussian distributions for each target. If the number of targets is T, then T separate Gaussian distributions are maintained. The number of Gaussian distributions is kept constant by integrating over the distribution of data associations of the previous step. This results in an algorithm where each of the target estimates gets updated by every measurement with weights that depend on the predicted probabilities of the associations. Gating is used for limiting the number of measurements for each track. If the predicted probabilities are too low (i.e., below a predefined threshold) for certain targets, those targets are not updated at all. Clutter measurements can be modeled similarly.

Sequential Monte Carlo (SMC) based multiple target tracking methods (Gordon, 1997; Karlsson and Gustafsson, 2001; Hue et al., 2001) typically belong to the class of *unique-neighbor data association* methods, as they are based on representing the data association and state posteriors as discrete sets of hypotheses.

These kind of SMC methods can be considered generalizations of MHT. Instead of maintaining the N most probable data association hypotheses, the joint tracking and data association problem is modeled as a Bayesian estimation problem and the *posterior distribution* is estimated with SMC methods. This *particle filtering* approach has the advantage that there are no restrictions on the analytic form of the model, although the required number of particles for a given accuracy can be high.

In article (Särkkä et al., 2004a) it is proposed how SMC based tracking and data association algorithms can be made more accurate and efficient using Rao-Blackwellization. In the Rao-Blackwellized Monte Carlo data association (RBMCDA) algorithm the states are integrated out in closed form and SMC is only used for the data association indicators. Instead of a pure particle representation, this leads to a mixture of Gaussians representation of the joint posterior distribution, which reduces variance and requires less particles for the same accuracy.

Approaches to Tracking Unknown Number of Targets

The JPDA method was originally formulated for a known number of targets, but it is possible to include track formation and termination logic in cascade with the algorithm (Bar-Shalom and Li, 1995). There is no explicit probability model for target appearance and disappearance, but instead there can be a Markov chain model for the number of data associations before track initiation is confirmed. Tracks are terminated when the probability of target existence goes below a predefined threshold.

MHT based methods (Bar-Shalom and Li, 1995; Blackman and Popoli, 1999; Stone et al., 1999) as well as PMHT (Streit and Luginbuhl, 1994) form hypotheses of associations with new targets for every measurement. In practice, to reduce the computational complexity, new target hypotheses (or new track hypotheses) are formed only when the measurement falls into an area where the likelihood of the association with the existing targets is too low. A track is deleted when its likelihood becomes too low compared to the other tracks.

Random sets and finite set statistics (FISST) (Mahler, 2004) provide a very general framework for Bayesian modeling of multiple target tracking in the case of an unknown number of targets. A tractable implementation of the framework is to use the first order moment of the multi-target posterior, the probability hypothesis density (PHD) (Mahler, 2003) as an approximation. SMC based implementations of the PHD have been reported, for example, in the articles (Vo et al., 2003; Sidenbladh, 2003). Vihola (2005) presents a random set particle filter based solution to tracking an unknown number of targets using bearings-only measurements.

In the SMC based method presented in the article (Hue et al., 2001) the extension to an unknown number of targets is based on hypothesis testing. Because the algorithm generates estimates of data association probabilities, these estimates can be used for approximating the probability of the hypothesis that the target has disappeared from the surveillance area. The detection of the appearance of a new target is based on testing the hypothesis between association with the old targets and with the new target.

The article (Kreucher et al., 2003) presents a SMC based method, which is similar to the method presented here except that a plain particle presentation of the joint posterior distribution is used. In the method, birth and death moves in particle proposals are used, and the moves resemble the birth and death models used here. Due to the plain particle presentation, the method in the article (Kreucher et al., 2003) is also applicable to the more general case of target tracking without explicit thresholding of measurements.

The method in (Doucet et al., 2002) also resembles the method presented here, except that the article does not suggest any particular form for the birth and death models. The approximation based on limiting the number of births and deaths on each time step is also discussed in (Doucet et al., 2002).

The particle filtering based method in (Isard and MacCormick, 2001) uses exponential (Poisson) models for target appearance and disappearance a bit similarly to the method here. The branching particle based solution (Ballantyne et al., 2001) also models target appearance as a stochastic (Markov) process.

The tracking of an unknown number of targets is also closely related to model selection. An application of SMC methods to estimating the number of RBF network weights from data is presented in (de Freitas et al., 2001).

In this section the SMC based RBMCDA method (Särkkä et al., 2004a) is extended to tracking an unknown number of targets. The extension is based on modeling the birth and death stochastic processes, such that track formation and termination (or initiation and deletion) are not based on heuristic rules, but on the rules determined by the estimation algorithm designed for the probability model. However, this probability model of the extension to an unknown number of targets is closer to the approach presented in (Stone et al., 1999) than to the approach in (Mahler, 2003) despite the more strict Bayesian nature of the latter.

4.1.2 RBMCDA with Known Number of Targets

Next the Rao-Blackwellized Monte Carlo data association (RBMCDA) method that was proposed in (Särkkä et al., 2004a) is reviewed and it is shown how it is related to the general Rao-Blackwellized particle filtering framework. The method described here is basically the same as in the original article except that the requirement of IID prior data association probabilities is replaced with a Markov chain assumption.

Filtering Model

The filtering model of the RBMCDA algorithm is the following:

- *Clutter or false alarm measurements* can be modeled using any probability density, which is independent of the target states $\mathbf{x}_k = (\mathbf{x}_{k,1} \ \cdots \ \mathbf{x}_{k,T})^T$

  $$p(\mathbf{y}_k \mid c_k = 0). \tag{4.1}$$

  Here c_k is the data association indicator, which has the value $c_k = 0$ for clutter and $c_k = j$ for the targets $j = 1, \ldots, T$. The clutter measurements can be, for example, uniformly distributed in the measurement space of volume V

  $$p(\mathbf{y}_k \mid c_k = 0) = 1/V. \tag{4.2}$$

- *Target measurements* are linear Gaussian

  $$p(\mathbf{y}_k \mid \mathbf{x}_{k,j}, c_k = j) = \mathrm{N}(\mathbf{y}_k \mid \mathbf{H}_{k,j}\mathbf{x}_{k,j}, \mathbf{R}_{k,j}). \tag{4.3}$$

  where the measurement matrices $\mathbf{H}_{k,j}$ and covariance matrices $\mathbf{R}_{k,j}$ can be different for each target. Non-linear measurement models can be used by replacing the non-linear model with a *locally linearized* model as in the extended Kalman filter (EKF) (Jazwinski, 1970; Bar-Shalom et al., 2001) or by using the *unscented transformation* as in the unscented Kalman filter (UKF) (Julier and Uhlmann, 2004b).

- *Target dynamics* are linear Gaussian

  $$p(\mathbf{x}_{k,j} \mid \mathbf{x}_{k-1,j}) = \mathrm{N}(\mathbf{x}_{k,j} \mid \mathbf{A}_{k-1,j}\mathbf{x}_{k-1,j}, \mathbf{Q}_{k-1,j}), \tag{4.4}$$

  where the transition matrix $\mathbf{A}_{k-1,j}$ and process noise covariance matrix $\mathbf{Q}_{k-1,j}$ may be different for different targets. The motions of individual targets are a priori independent. Because for any continuous time linear Gaussian dynamic model there exists an equivalent discrete time linear Gaussian model (Jazwinski, 1970; Bar-Shalom et al., 2001), this directly generalizes to continuous time linear models. Non-linear discrete or continuous time dynamic models can also be used by forming Gaussian approximations by EKF (Jazwinski, 1970; Bar-Shalom et al., 2001) or by UKF (Julier and Uhlmann, 2004b). With slight modifications to the algorithm, it would be possible to use the interacting multiple model (IMM) filter (Bar-Shalom et al., 2001), which allows the modeling of target maneuvers.

- *Target and clutter association priors* are known and can be modeled as an mth order Markov chain

  $$p(c_k \mid c_{k-1}, \ldots, c_{k-m}). \tag{4.5}$$

This is kind of higher order models are needed for implementing constraints to the data associations on a single measurement scan. Of course, the model can also be a first order Markov chain ($m = 1$) or the associations can be completely independent ($m = 0$).

- *The number of targets* T is known and constant.
- *Target state priors* can be represented as a weighted importance sample set

$$p(\mathbf{x}_{0,j}) = \sum_i w^{(i)} \, \mathrm{N}(\mathbf{x}_{0,j} \mid \mathbf{m}_{0,j}^{(i)}, \mathbf{P}_{0,j}^{(i)}). \tag{4.6}$$

Relationship with the Rao-Blackwellized Particle Filter

The RBMCDA model described above fits exactly to the Rao-Blackwellized particle filtering framework (see, e.g., Doucet et al., 2001)) when the sampled latent variable $\boldsymbol{\lambda}_k$ is defined to contain the data association event indicators, c_k at time step k

$$\boldsymbol{\lambda}_k = c_k. \tag{4.7}$$

1. The state $\mathbf{x}_k$ on time step k consists of the stacked vector of target states

$$\mathbf{x}_k = \begin{pmatrix} \mathbf{x}_{k,1} \\ \vdots \\ \mathbf{x}_{k,T} \end{pmatrix}. \tag{4.8}$$

2. The prior distribution of the joint state $p(\mathbf{x}_0)$ is Gaussian, because the individual target prior distributions are Gaussian.

3. The joint dynamic model of targets is linear Gaussian

$$p(\mathbf{x}_k \mid \mathbf{x}_{k-1}) = \mathrm{N}(\mathbf{x}_k \mid \mathbf{A}_{k-1}\mathbf{x}_{k-1}, \mathbf{Q}_{k-1}), \tag{4.9}$$

where $\mathbf{A}_{k-1}$ is a block diagonal matrix consisting of the dynamic models of the targets and $\mathbf{Q}_{k-1}$ is the block diagonal process noise covariance.

4. The joint measurement model of the targets can be written as

$$p(\mathbf{y}_k \mid \mathbf{x}_k, c_k) = \mathrm{N}(\mathbf{y}_k \mid \mathbf{H}_k(c_k)\mathbf{x}_k, \mathbf{R}_k(c_k)), \tag{4.10}$$

where the measurement model matrix $\mathbf{H}_k(c_k)$ is formed conditional on the data association c_k such that the only nonzero entries are the ones corresponding to the measurement model of the target c_k. The measurement noise covariance $\mathbf{R}_k(c_k)$ is the measurement noise covariance of the target c_k.

5. The clutter measurements are simply state independent measurements with a measurement model of the form $p(\mathbf{y}_k \mid c_k = 0)$.
6. The data associations are modeled as a Markov chain of latent variables as in the Rao-Blackwellized particle filter model.

Sequential Measurement Update Issues

Up to now we have formulated the model such that exactly one measurement is obtained at one time step t_k. This way, the data association procedure can be reduced to processing one measurement at a time. However, this does not mean that we are restricted to one measurement per time instance (i.e., scan), because the successive time steps may occur at the same time instance.

Table 4.1: Several measurement steps can occur on the same time instance or scan.

scan 1 (3 measurements):	$t_1 = t_2 = t_3 = 1$ s
scan 2 (2 measurements):	$t_4 = t_5 = 2$ s
scan 3 (3 measurements):	$t_6 = t_7 = t_8 = 2.5$ s
scan 4 (1 measurement):	$t_9 = 3.5$ s
$\cdots$	$\cdots$

In the example presented in Table 4.1 the time steps 1,2 and 3 all occur on time instance 1 s, time steps 4 and 5 occur on time instance 2 s and so on. Now, for example, time steps 1,2 and 3 can be processed sequentially by setting $\Delta t = 0$ in the dynamic model and in the prediction step, which is equivalent to not performing the prediction step at all between the measurements.

The original RBMCDA that was presented in (Särkkä et al., 2004a) had the restriction that the data associations were required to be conditionally independent as in PMHT (Streit and Luginbuhl, 1994). This makes it impossible to model certain joint effects in sets of measurements obtained at one time instance (i.e., scan). For example, it is impossible to restrict the maximum number of data associations with each target to one per time instance. However, by allowing the data association priors to depend on previous data associations (at least on the same time instance) this restriction can be included in the model.

As already discussed in this section, the data association indicators may also depend on previous data associations, that is, they may form an mth order Markov chain

$$p(c_k \mid c_{k-1}, \ldots, c_{k-m}). \tag{4.11}$$

This kind of model can be used for restricting the data associations to at most one data association to each target at single time instance as follows:

- The joint prior model can be written in the general form

$$p(c_{k+m-1}, \dots, c_k). \tag{4.12}$$

  We can encode the desired restriction into the prior by assigning zero probability to any joint event, which would have two associations to the same target.

- The joint model can also be expanded as follows:

$$\begin{aligned} & p(c_{k+m-1}, \dots, c_k) \\ & = \prod_{j=1}^{m} p(c_{k+j} \mid c_k, \dots, c_{k+j-1}). \end{aligned} \tag{4.13}$$

  This means that exactly the same model can be realized by using using the following priors for the data associations:

  - c_k has the prior $p(c_k)$.
  - c_{k+1} has the prior $p(c_{k+1} \mid c_k)$.
  - $\dots$
  - c_{k+m-1} has the prior $p(c_{k+m-1} \mid c_{k+m-2}, \dots, c_k)$.

Using the above idea it is possible to reduce any joint data association prior of form (4.12) to an equivalent Markov prior model, which is suitable for sequential processing.

Assume, for example, that we are tracking two targets and on each time instance we may obtain zero or one detections from each of the targets. The rest of the detections are false alarms, that is, clutter. The target detection probabilities of both targets are the same and given as

$$p(\text{detection}) = p_d. \tag{4.14}$$

Assume that on time instance t_{scan} we obtain m measurements $\mathbf{y}_k, \dots, \mathbf{y}_{k+m-1}$ (i.e., the times of the steps are $t_k, ..., t_{k+m-1} = t_{\text{scan}}$). The prior for the data associations can be now defined sequentially as follows:

- Define detection indicators as follows

$$\begin{aligned} \delta_1(j) &= \begin{cases} 1, & \text{if there is target 1 detection in } c_{k+j-1} ... c_k \\ 0, & \text{otherwise} \end{cases} \\ \delta_2(j) &= \begin{cases} 1, & \text{if there is target 2 detection in } c_{k+j-1} ... c_k \\ 0, & \text{otherwise}. \end{cases} \end{aligned} \tag{4.15}$$

- Compute the data association priors given each of the possible detection indicator combinations:

$$\begin{aligned} &p(c_{k+j} \mid \delta_1(j) = 0 \text{ and } \delta_2(j) = 0) \\ &p(c_{k+j} \mid \delta_1(j) = 0 \text{ and } \delta_2(j) = 1) \\ &p(c_{k+j} \mid \delta_1(j) = 1 \text{ and } \delta_2(j) = 0) \\ &p(c_{k+j} \mid \delta_1(j) = 1 \text{ and } \delta_2(j) = 1). \end{aligned} \tag{4.16}$$

Prior models having the restriction of one data association to each target per time instance can be defined in similar manner for any number of targets. The resulting prior is of the recursive (Markov) form, which is a special case of the model described in this section.

Data Representation

The algorithm state consists of a set of N particles, where each particle i at time step k contains the following:

$$\{c^{(i)}_{k-m+1:k}, \mathbf{m}^{(i)}_{k,1}, \ldots, \mathbf{m}^{(i)}_{k,j}, \ldots, \mathbf{m}^{(i)}_{k,T}, \mathbf{P}^{(i)}_{k,1}, \ldots, \mathbf{P}^{(i)}_{k,j}, \ldots, \mathbf{P}^{(i)}_{k,T}, w^{(i)}_k\}, \tag{4.17}$$

where

- $c^{(i)}_{k-m+1:k}$ are the data association indicators of time steps $k - m + 1, \ldots, k$ with integer values $0, \ldots, T$, where T is the number of targets. If the data association prior model is an mth order Markov model, then m previous data associations should be stored. If the data association prior is time independent, the data association indicators do not need to be stored at all.
- $\mathbf{m}^{(i)}_{k,j}$,$\mathbf{P}^{(i)}_{k,j}$ are the mean and covariance of the target j, and they are conditional on the data association history $c^{(i)}_{1:k}$.
- $w^{(i)}_k$ is the importance weight of the particle.

Evaluating and Sampling from the Optimal Importance Distribution

For each particle i, the optimal importance distribution is given by

$$p(c_k \mid \mathbf{y}_{1:k}, c^{(i)}_{1:k-1}). \tag{4.18}$$

The marginal measurement likelihood is given by (Särkkä et al., 2004a)

$$\begin{aligned} &p(\mathbf{y}_k \mid c_k, \mathbf{y}_{1:k-1}, c^{(i)}_{1:k-1}) \\ &\quad = \begin{cases} 1/V & \text{if } c_k = 0 \\ \mathrm{KF}_{lh}(\mathbf{y}_k, \mathbf{m}^{-(i)}_{j,k}, \mathbf{P}^{-(i)}_{j,k}, \mathbf{H}_{j,k}, \mathbf{R}_{j,k}) & \text{if } c_k = j \end{cases} \end{aligned} \tag{4.19}$$

where $j = 1, \ldots, T$ and $\mathrm{KF}_{lh}(\cdot)$ denotes the Kalman filter measurement likelihood evaluation. $\mathbf{H}_{j,k}$ and $\mathbf{R}_{j,k}$ are the measurement model matrix and the measurement covariance matrix of the target j, respectively. For $j = 1, \ldots, T$ we have

$$[\mathbf{m}_{j,k}^{-(i)}, \mathbf{P}_{j,k}^{-(i)}] = \mathrm{KF}_p(\mathbf{m}_{j,k-1}^{(i)}, \mathbf{P}_{j,k-1}^{(i)}, \mathbf{A}_{j,k-1}, \mathbf{Q}_{j,k-1}), \tag{4.20}$$

where $\mathrm{KF}_p(\cdot)$ denotes the Kalman filter prediction step, and $\mathbf{m}_{j,k-1}^{(i)}, \mathbf{P}_{j,k-1}^{(i)}$ are the mean and the covariance of target j in particle i, which is conditioned on the state history $c_{1:k-1}^{(i)}$. $\mathbf{A}_{j,k-1}$ and $\mathbf{Q}_{j,k-1}$ are the transition matrix of dynamic model and the process noise covariance matrix of the target j, respectively.

The posterior distribution of c_k can be calculated using Bayes' rule

$$\begin{aligned} &p(c_k \mid \mathbf{y}_{1:k}, c_{1:k-1}^{(i)}) \\ &\qquad \propto p(\mathbf{y}_k \mid c_k, \mathbf{y}_{1:k-1}, c_{1:k-1}^{(i)}) \\ &\qquad \times p(c_k \mid c_{k-m:k-1}^{(i)}), \end{aligned} \tag{4.21}$$

where we have used the fact that an association c_k does not depend on the previous measurements $\mathbf{y}_{1:k-1}$, and depends only on the m previous associations $c_{k-m:k-1}$ if the order of the Markov model is m.

We can sample from the optimal importance distribution as follows:

1. Compute the unnormalized clutter association probability

$$\begin{aligned} \hat{\pi}_0^{(i)} &= p(\mathbf{y}_k \mid c_k^{(i)} = 0, \mathbf{y}_{1:k-1}, c_{1:k-1}^{(i)}) \\ &\times p(c_k^{(i)} = 0 \mid c_{k-m:k-1}^{(i)}). \end{aligned} \tag{4.22}$$

2. Compute the unnormalized target association probabilities for each target $j = 1, \ldots, T$

$$\begin{aligned} \hat{\pi}_j^{(i)} &= p(\mathbf{y}_k \mid c_k^{(i)} = j, \mathbf{y}_{1:k-1}, c_{1:k-1}^{(i)}) \\ &\times p(c_k^{(i)} = j \mid c_{k-m:k-1}^{(i)}). \end{aligned} \tag{4.23}$$

3. Normalize the importance distribution:

$$\pi_j^{(i)} = \frac{\hat{\pi}_j^{(i)}}{\sum_{j'=0}^{T} \hat{\pi}_{j'}^{(i)}}, \quad j = 0, \ldots, T. \tag{4.24}$$

4. Sample a new association $c_k^{(i)}$ with the following probabilities:

   - Draw $c_k^{(i)} = 0$ with probability $\pi_0^{(i)}$

- Draw $c_k^{(i)} = 1$ with probability $\pi_1^{(i)}$.
- Draw $c_k^{(i)} = 2$ with probability $\pi_2^{(i)}$.
- ...
- Draw $c_k^{(i)} = T$ with probability $\pi_T^{(i)}$.

Now it is easy to evaluate also the probabilities of data association hypotheses in the optimal importance distribution, because the probabilities are given by the terms $\pi_j^{(i)}, j = 1, \ldots, T$ conditional on each particle.

Algorithm Implementation

As already shown in this section the RBMCDA algorithm is a special case of the generic RBPF algorithm. However, due to the conditional independences between the targets, the full Kalman filter prediction and update steps for all targets can be reduced to single target predictions and updates. Actually, because the targets are a priori independent, conditional on the data associations c_k the targets will remain independent during tracking. This leads to the following simplifications to the RBPF computations:

1. The Kalman filter prediction steps can be done for each target in each particle separately. That is, we do not need to do Kalman filter prediction to the joint mean and covariance of all targets, but only to each target separately.

2. We can always use the optimal importance distribution, which was described in the previous section, as the importance distribution for the latent variables.

3. The marginalized measurement likelihoods can be computed for each target separately. Note that these likelihoods have already been computed for each target during the evaluation of the optimal importance distribution.

4. The measurement updates can also be performed for each target separately. This means that the Kalman filter update is actually performed only to one target in each particle.

4.1.3 RBMCDA with Unknown Number of Targets

In the next sections we extend the RBMCDA algorithm to an unknown and time varying number of targets. The probabilistic construction of the model for an unknown number of targets, that is, the probability model for varying state space dimension follows roughly the approach described in (Stone et al., 1999), which extends MHT to tracking an unknown number of targets. The idea is to assume that there is always a (very large) *constant number of targets* T_∞. But *an unknown, varying number of them are visible (or alive)*, and they are the ones we are tracking. The visibility of targets is represented with a vector of discrete indicator variables.

The number T_∞ ensures that the joint distribution of the target states is a Gaussian distribution with constant dimensionality and thus a well defined probability distribution. The model is formulated such that we do not explicitly need to know the actual number of targets T_∞ as long as it is theoretically large enough.

As an extension to the MHT method, we construct a probabilistic stochastic process model for the births and deaths of the targets and show how Rao-Blackwellized particle filtering can be applied to this problem.

Filtering Model

1. *Clutter or false alarm measurements* have the same kind of model as in the RBMCDA model in Section 4.1.2.

2. *Target measurements* are (approximately) linear Gaussian as in the RBMCDA model in Section 4.1.2. EKF and UKF based approximations can be used for handling non-linear models.

3. *Target dynamics* are linear (approximately) Gaussian as in the RBMCDA model in Section 4.1.2. EKF, UKF or IMM can be used for handling more general dynamic models.

4. The target ($c_k = j$) and clutter ($c_k = 0$) association priors, in the case that births and deaths do not occur at the current time step, are known and can be modeled as an mth order Markov chain $p(c_k \mid c_{k-m:k-1}, T_{k-m:k-1})$, where $T_{k-m:k-1}$ contains the number of targets at time steps $k-m,\ldots,k-1$. For example, we might have a uniform prior over targets and clutter:

$$p(c_k \mid c_{k-m:k-1}, T_{k-m:k-1}) = \frac{1}{1+T_{k-1}}. \tag{4.25}$$

5. Target births may happen only when a measurement is obtained, and in that case a birth happens with probability p_b. For simplicity, the model is defined such that a birth may happen only jointly with an association event, so that if there is no association to a newborn target, there is no birth.

   This is equivalent to stating that the target state prior remains constant until the first measurement is associated, that is, the dynamic model does not affect the target state before the first measurement has been associated to the target. This indicates that it is sufficient to consider the time of the first associated measurement as the actual birth moment.

6. After associating a measurement with a target, the life time t_d (or time to death) of the target has a known probability density
$$t_d \sim p(t_d), \tag{4.26}$$
which can be, for example, an exponential or gamma distribution.
7. At the time of birth each target has a known Gaussian prior distribution
$$p(\mathbf{x}_{k_0,j}) = \mathrm{N}(\mathbf{x}_{0,j} \mid \mathbf{m}_{0,j}, \mathbf{P}_{k_0,j}). \tag{4.27}$$

Probabilities of Birth and Death

If a birth has occurred, it is assumed to be certain that the current measurement is associated to the newborn target:
$$p(c_k \mid \text{birth}) = \begin{cases} 1 & , \quad \text{if } c_k = T_{k-1} + 1 \\ 0 & , \quad \text{otherwise}. \end{cases} \tag{4.28}$$

In the case of no birth, the Markov model for the data associations applies:
$$p(c_k \mid \text{no birth}) = p(c_k \mid c_{k-m:k-1}). \tag{4.29}$$

The data association and birth events can be divided into the following cases with different probabilities:

1. A target is born and the measurement is associated with the newborn target:
$$\begin{aligned} b_k &= \text{birth} \\ c_k &= T_{k-1} + 1. \end{aligned} \tag{4.30}$$
2. A target is not born and the measurement is associated with one of the existing targets or with clutter:
$$\begin{aligned} b_k &= \text{no birth} \\ c_k &= j, \qquad j = 0, \ldots, T_k. \end{aligned} \tag{4.31}$$
3. Other events have zero probability.

Thus, given the associations $c_{k-m:k-1}$ on the m previous steps, the joint distribution of the event $b_k \in \{\text{no birth}, \text{birth}\}$ and the association c_k is given as
$$\begin{aligned} & p(b_k, c_k \mid c_{k-m:k-1}) = \\ & \qquad \begin{cases} p_b & \text{in case (1)} \\ (1 - p_b)\, p(c_k \mid c_{k-m:k-1}) & \text{in case (2)} \\ 0 & \text{in case (3),} \end{cases} \end{aligned} \tag{4.32}$$

where p_b is the prior probability of birth.

The restriction to one data association per target at single time instance in the case of unknown number of targets can be handled in the same manner as in the case of known number of targets. We simply assume that there is positive probability of detecting a newborn target on each step. That is, in addition to the existing targets we model the possibility detection of a new target which has the detection probability p_b. This probability of detecting a new target is equivalent to the probability of birth, because we have defined the birth to be the event of detecting the target for the first time. However, the difference to the restriction of one association per target is that births may occur as many times as there are measurements on scan, not only once per scan.

The filtering model presented in this section states that after associating a measurement with a target, the life time t_d of the target has the known probability density (4.26). Thus if the last association with target j was at time $\tau_{k,j}$, and on the previous time step t_{k-1} we sampled a hypothesis that the target is alive, then the probability that the target is dead at current time step t_k is

$$\begin{aligned} & p(\text{death of } j \mid t_k, t_{k-1}, \tau_{k,j}) \\ & \quad = P(t_d \in [t_{k-1} - \tau_{k,j}, t_k - \tau_{k,j}] \mid t_d \geq t_{k-1} - \tau_{k,j}). \end{aligned} \tag{4.33}$$

Relationship to RBPF

The RBMCDA algorithm with an unknown number of targets fits to the RBPF framework, if the latent variable $\boldsymbol{\lambda}_k$ contains the visibility indicator $\mathbf{e}_k$ and the data association indicator c_k at the current time step

$$\boldsymbol{\lambda}_k = \{\mathbf{e}_k, c_k\}. \tag{4.34}$$

The visibility indicators and the data associations implicitly define the number of (visible) targets T_k at each time step.

Given that the targets are a priori unordered, there is a high permutation symmetry in the posterior distributions of the target states, visibility indicators and data association indicators. We can change the indices of any two targets, including the visibility indicators and data associations, and the probability of the configuration will remain the same. For this reason, we shall select one of the permutations arbitrarily and use it for representing all the permutations. This permutation is based on the times of the first associations with the targets. This does not change the model, because this is not a priori ordering, but is merely a way of selecting a compact representation for a very high number of redundant permutations.

1. The joint state $\mathbf{x}_k$ contains the states of the T_∞ targets
$$\mathbf{x}_k = \begin{pmatrix} \mathbf{x}_{k,1} \\ \ldots \\ \mathbf{x}_{k,T_\infty} \end{pmatrix}. \tag{4.35}$$

2. At the initial time step the targets have Gaussian prior distributions
$$\mathrm{N}(\mathbf{x}_{0,j} \mid \mathbf{m}_0, \mathbf{P}_0). \tag{4.36}$$

   The model is constructed such that the invisible targets at any time step k (indicated by $\mathbf{e}_k$) do not have a dynamic model. This means that the targets which have not yet become visible (have not been born yet) at any time step k have independent Gaussian prior distributions $\mathrm{N}(\mathbf{x}_{k,j} \mid \mathbf{m}_0, \mathbf{P}_0)$. If we denote the sets of not visible and visible target indices with $\mathcal{J}_0$ and $\mathcal{J}_1$, respectively, the joint prior distribution of all targets is of the form
$$\begin{aligned} p(\mathbf{x}_k \mid \mathbf{y}_{1:k}) = & \prod_{j \in \mathcal{J}_1} \mathrm{N}(\mathbf{x}_{k,j} \mid \mathbf{m}_{k,j}, \mathbf{P}_{k,j}) \\ & \times \prod_{j' \in \mathcal{J}_0} \mathrm{N}(\mathbf{x}_{k,j'} \mid \mathbf{m}_0, \mathbf{P}_0). \end{aligned} \tag{4.37}$$

   That is, the distribution of the visible targets is completely independent of the distributions of the invisible targets and thus it suffices to store only the states of the visible targets instead of all T_∞ targets. The joint distribution of all targets is still always theoretically Gaussian distribution of dimension T_∞.

3. When a target birth occurs, that is, a new target becomes visible (i.e., produces the first measurement) a new item in the indicator vector $\mathbf{e}_k$ is set and the corresponding target prior distribution is updated (initialized) by the measurement. Because we only need to store one possible permutation from a high number of equivalent target permutations, we can add the new target to the first empty place in the indicator vector $\mathbf{e}_k$.

4. When a target dies, that is, becomes invisible again, the target distribution again becomes the prior and the target state is moved to the end of the joint state vector $\mathbf{x}_k$ and indicator vector $\mathbf{e}_k$. The targets in the vectors can be shifted such that the visible targets always remain in the beginning of the vectors.

5. The target dynamics, target measurements and clutter measurements are modeled in the same way as in the RBMCDA model of Section 4.1.2 and thus they fit into the RBPF framework easily.

6. By constructing a prior model for births and deaths, we get the distribution

$$p(\mathbf{e}_k \mid \mathbf{e}_{k-1}), \tag{4.38}$$

which defines the dynamics of births and deaths. The data association model is of the form

$$p(c_k \mid c_{k-m:k-1}, \mathbf{e}_k), \tag{4.39}$$

and thus these two models together give a joint Markov chain model for the indicators:

$$\begin{aligned} & p(\mathbf{e}_k, c_k \mid c_{k-m:k-1}, \mathbf{e}_{k-m:k-1}) \\ & = p(c_k \mid c_{k-m:k-1}, \mathbf{e}_k)\, p(\mathbf{e}_k \mid \mathbf{e}_{k-1}), \end{aligned} \tag{4.40}$$

which is the form required by the RBPF model.

Evaluating and Sampling from the Optimal Importance Distribution

The possible events between two measurements $\mathbf{y}_{k-1}$ and $\mathbf{y}_k$ and at the association of measurement $\mathbf{y}_k$ are:

1. Targets may die (indicated by elements of $\mathbf{e}_k$):
   (a) none of the targets dies
   (b) one or more targets die
2. $\mathbf{y}_k$ is associated with (indicated by c_k):
   (a) clutter
   (b) one of the existing targets
   (c) a newborn target

Death events are independent of the measurements. However, the two event families are related such that a new measurement $\mathbf{y}_k$ can be associated only to the targets that have not died between the measurements $\mathbf{y}_{k-1}$ and $\mathbf{y}_k$.

The model (4.40) assigns unique prior probabilities to each of the finite number of different events, but the problem is that the number of possible events grows exponentially with the number of targets. The combinatorial problem in the number of events is solely due to the exponential number of possible combinations of target deaths. There is no combinatorial problem in target births, because we can always use the sequential update scheme as discussed in Section 4.1.2. However, the purpose of the death model is only to remove the targets with which no measurements have been associated for a long time. Because the death model is built only for serving this purpose without any physical meaning, it is not desirable to spend most of the computing power on it.

For the above reason we shall use an approximation where we restrict the possible number of deaths at each time step to one. This means that at each time step (measurement time) either none or exactly one of the targets dies. The probability of two or more death events between measurements is assumed to be zero. Note that the varying time step size is accounted already in the death model (4.33) and thus the survival rate will change only a bit, because we assume that there cannot be more deaths than there are measurements. We could overcome this restriction by performing redundant measurement update steps, for example, by generating artificial clutter measurements, which have prior probability one of being clutter.

Because due to restricting the number of deaths on each time step the number of events grows only linearly with the number of targets, we can construct an approximate prior distribution of births, deaths and associations as follows:

1. Enumerate all possible combinations of joint birth, (zero or one) deaths, and association events and compute probabilities for each of the combinations.
2. Normalize the list of events such that their probabilities sum to one.

For each combination of birth, death and association events there is a transition pair $(\mathbf{e}_{1:k-1}, c_{1:k-1}) \to (\mathbf{e}_k, c_k)$ with a probability given by the above procedure. That is, we have an approximate representation of the distribution

$$p(\mathbf{e}_k, c_k \mid \mathbf{e}_{1:k-1}, c_{1:k-1}). \tag{4.41}$$

The likelihood term $p(\mathbf{y}_k \mid \mathbf{e}_k, c_k)$ can be computed similarly as in the case of a known number of targets (see Equation (4.19)). By multiplying each of the birth, death, and association combinations with the measurement likelihood and normalizing, we can form the optimal importance distribution similarly as in Section 4.1.2.

Data Representation

The algorithm state consists of a set of N particles, where each particle i at time step k contains the following:

$$\{c^{(i)}_{k-m+1:k}, \mathbf{e}^{(i)}_k, \mathbf{m}^{(i)}_{k,1}, \ldots, \mathbf{m}^{(i)}_{k,j}, \ldots, \mathbf{m}^{(i)}_{k,T}, \mathbf{P}^{(i)}_{k,1}, \ldots, \mathbf{P}^{(i)}_{k,j}, \ldots, \mathbf{P}^{(i)}_{k,T}, w^{(i)}_k\}, \tag{4.42}$$

where

- $c^{(i)}_{k-m+1:k}$ are the data association indicators of the time steps $k-m+1, \ldots, k$.
- $\mathbf{e}^{(i)}_k$ is the life-indicator, which is a binary vector of length T_∞ indicating which of the targets are alive at current time step.

- $\mathbf{m}_{k,j}^{(i)}$,$\mathbf{P}_{k,j}^{(i)}$ are the mean and covariance of the target j, and they are conditional on the data association history $c_{1:k}^{(i)}$.
- $w_k^{(i)}$ is the importance weight of the particle.

The following information is also implicitly or explicitly stored for each particle:

$$\{T_k^{(i)}, \tau_{k,j}^{(i)}, \mathrm{id}_{k,j}^{(i)}, \}, \tag{4.43}$$

where

- $T_k^{(i)}$ is the number of targets.
- $\tau_{k,j}^{(i)}$ is the time of the last measurement associated with target j.
- $\mathrm{id}_{k,j}^{(i)}$ is a unique integer valued identifier, unique over all targets in all particles, which is assigned at the birth of the target.

Algorithm Implementation

Similarly to the case of RBMCDA with known number of targets, because the targets are a priori independent, conditional on data associations c_k and indicators $\mathbf{e}_k$, the targets will also remain independent during tracking. This means that exactly the same simplifications to RBPF apply to the case of an unknown number of targets as to a known number of targets.

4.1.4 Simulations

Bearings Only Tracking of a Known Number of Targets

First we shall consider a classical bearings only multiple target tracking problem, which frequently arises in the context of passive sensor tracking. The simulation scenario is similar to that was presented in (Särkkä et al., 2004a), but now the simulation includes clutter measurements and the restriction of one data association per target on single time instance is also modeled.

There are two targets on the scene and the dynamics of target j with the state vector $\mathbf{x}_{j,k} = (x_{j,k}\ y_{j,k}\ \dot{x}_{j,k}\ \dot{y}_{j,k})^T$ can be modeled with a discretized Wiener velocity model (Bar-Shalom et al., 2001)

$$\begin{pmatrix} x_{j,k} \\ y_{j,k} \\ \dot{x}_{j,k} \\ \dot{y}_{j,k} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \Delta t & 0 \\ 0 & 1 & 0 & \Delta t \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_{j,k-1} \\ y_{j,k-1} \\ \dot{x}_{j,k-1} \\ \dot{y}_{j,k-1} \end{pmatrix} + \mathbf{q}_{k-1}, \tag{4.44}$$

where $\mathbf{q}_{k-1}$ is the Gaussian process noise with moments

$$\begin{aligned}\mathrm{E}[\mathbf{q}_{k-1}] &= \mathbf{0} \\ \mathrm{E}[\mathbf{q}_{k-1}\mathbf{q}_{k-1}^T] &= \begin{pmatrix} \frac{1}{3}\Delta t^3 & 0 & \frac{1}{2}\Delta t^2 & 0 \\ 0 & \frac{1}{3}\Delta t^3 & 0 & \frac{1}{2}\Delta t^2 \\ \frac{1}{2}\Delta t^2 & 0 & \Delta t & 0 \\ 0 & \frac{1}{2}\Delta t^2 & 0 & \Delta t \end{pmatrix} q,\end{aligned}$$

where q is the spectral density of the noise. In the simulation, the value $q = 0.1$ was used for both the targets. The noise in an angular measurement from target j by sensor i can be modeled as Gaussian

$$\theta_k = \arctan\left(\frac{y_{j,k} - s_y^i}{x_{j,k} - s_x^i}\right) + r_k, \tag{4.45}$$

where (s_x^i, s_y^i) is the position of sensor i and $r_k \sim \mathrm{N}(0, \sigma^2)$ with $\sigma = 0.02$ radians.

Because the measurement model is non-linear we replace the Kalman filter in the data association algorithm with EKF. The uncertainty in data associations can be modeled by defining a variable c_k, which has the value $c_k = j$ if the measurement at time step k is associated with target j.

The target detection probability is set to $p_d = 80\%$ and the number of clutter measurements at a single time instance (uniform on range $[-\pi, \pi]$) is Poisson-distributed with mean 5. The measurement data obtained from simulated sensors is shown in Figure 4.4. The initial distribution was on purpose selected such that all the four crossings of measurements from the two sensors contain some probability mass, and the distributions of the targets are two-modal as shown in Figure 4.5. The particles in the figure are a random sample drawn from the posterior distribution estimate, used for visualizing the distribution. The actual posterior distribution estimate is a mixture of Gaussians which is hard to visualize directly. The number of Monte Carlo samples used in the estimation method was 100.

Figure 4.6 shows the final tracking result, and it can be seen that in the beginning of the trajectory the posterior distribution is multi-modal. Also the posterior distributions of the trajectories are a bit wider in the area between the sensors, because in that area the position uncertainty is higher in the direction of the line connecting the two angular sensors. Again, particles are used for visualizing the distribution, although the true posterior distribution estimate is a mixture of Gaussians.

Figure 4.7 shows the *smoothed* tracking result, which is an estimate where the distributions of all time steps are conditioned on all the measurements. This kind of estimate can be easily calculated with (fixed interval) Kalman smoothers

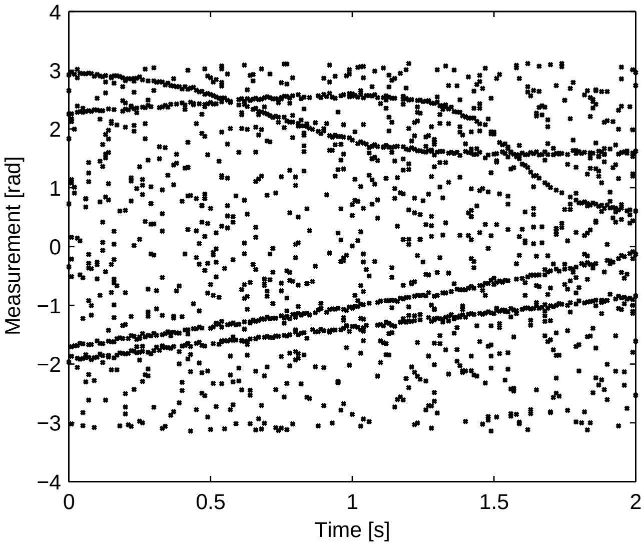

Figure 4.4: Measurement data obtained from angular sensors.

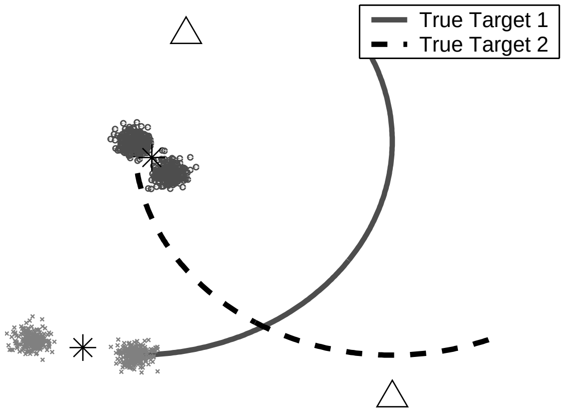

Figure 4.5: The prior distributions of the targets. Half of the prior probability mass is located in the wrong sensor measurement crossings.

(Jazwinski, 1970; Bar-Shalom et al., 2001) and particle smoothers (Kitagawa, 1996) also in the Rao-Blackwellized particle filtering case. Conditional on all the measurements the trajectory no longer contains multi-modalities. Also the position uncertainty on the line connecting the sensors is lower.

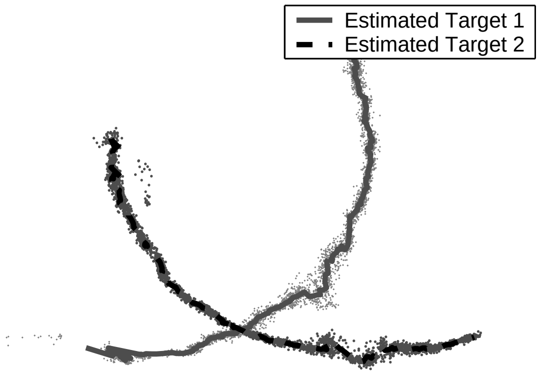

Figure 4.6: Filter estimates for each time step. In the beginning of the trajectory the posterior distribution is multi-modal. The multivalued can be seen from the two extra clouds of particles, which are located quite far away from the actual target trajectories. The higher position uncertainty on the line connecting the two sensors can also be seen.

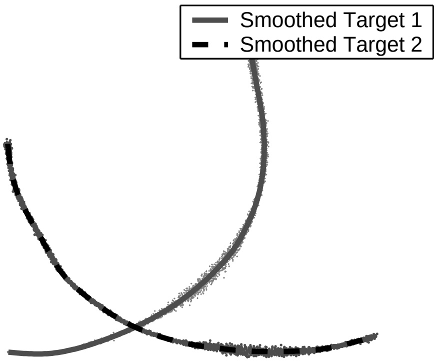

Figure 4.7: Smoothed estimates do not have the multivalued in the beginning of the trajectory, because later measurements have resolved it. The position uncertainty on the line connecting the two sensors is also lower than in the filter estimates.

Unknown Number of 1D Signals

The algorithm extension to an unknown number of signals was simulated with the appearing and disappearing signals as described in Table 4.2. The true signals and simulated data are shown in the Figure 4.8. All the signals are modeled with discretized white noise acceleration models (Bar-Shalom et al., 2001)

$$\begin{pmatrix} x_k \\ \dot{x}_k \end{pmatrix} = \begin{pmatrix} 1 & \Delta t \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_{k-1} \\ \dot{x}_{k-1} \end{pmatrix} + \mathbf{q}_{k-1}, \tag{4.46}$$

Table 4.2: Signal appearance and disappearance schedule in the 1D scenario with an unknown number of signals.

Signal	*Appears*	*Disappears*
$x^{(1)}(t)$	$t = 0$	$t = 8$
$x^{(2)}(t)$	$t = 0$	$t = 15$
$x^{(3)}(t)$	$t = 1$	$t = 4$
$x^{(4)}(t)$	$t = 2$	$t = 5$
$x^{(5)}(t)$	$t = 5.5$	$t = 10$
$x^{(6)}(t)$	$t = 6$	$t = 15$

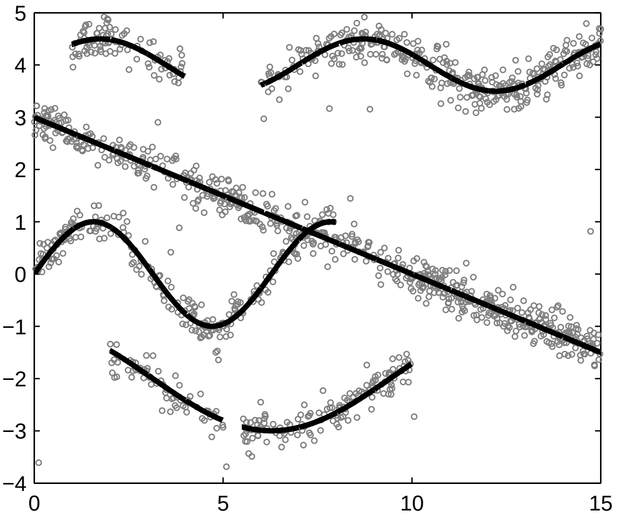

Figure 4.8: Simulated data of the 1D scenario with an unknown number of signals.

where $x_k = x(t_k), \dot{x}_k = \dot{x}(t_k)$, the sampling period is $\Delta t = 1/100$, and the process noise spectral density is $q = 1/10$. The signal conditioned measurements are modeled (and simulated) as measurements of the signal plus a white Gaussian noise component

$$y_{k,j} = x^{(j)}(t_k) + r_k, \tag{4.47}$$

where $r_k \in \mathrm{N}(0, 1/5^2)$, given that the measurement is from signal j. Every measurement has an equal chance of originating from each of the visible signals and 1% change of being a corrupted measurement uniformly distributed on the area $[-5, 5]$. The prior distribution for a new born signal was a Gaussian distribution with mean $\mathbf{m}_0 = \begin{pmatrix} 0 & 0 \end{pmatrix}^T$ and covariance $\mathbf{P}_0 = \mathrm{diag}(100, 10)$. The number of signals is unknown and the following model for the births and deaths is used:

- The prior probability of birth $p_b = 1/100$.
- A priori time to death t_d from the last data association has the gamma distribution[1]

  $$t_d \sim \mathrm{Gamma}(t_d \mid \alpha, \beta), \tag{4.48}$$

  with constant parameters α and β.

The number of Monte Carlo samples used in estimation was $N = 10$. Figure 4.9 shows the result of filtering with parameters $\alpha = 2$ and $\beta = 1$. The plotted result is the hypothesis contained in the particle with largest weight. As can be seen from Figure 4.10 there is slight delay after the disappearance of the signals before they disappear from estimation. The longest delay is in the signal that ends very near the other signal. Also the 1 time step gap in the lowest signal is not detected. Figure 4.11 shows the result of applying a Kalman smoother to the filtering result in the particle with the largest weight. This corresponds to the maximum a posteriori signal estimate. It can be seen that the estimation result follows the actual signal paths quite well except for the slight delays in signal disappearance.

Figures 4.12, 4.13 and 4.14 show the results when the disappearance model parameters were set to $\alpha = 2$ and $\beta = 10$, which means that the signals disappear almost ten times faster than in the previous figures. It can be seen that in this case the signals are estimated to disappear and reappear also when there are random gaps in the signals due to uneven measurement times. This result is quite natural, because the model states that even quite small gaps in the measurement sequence (or actually in the signal) should be interpreted as disappearances and reappearances of the signal.

Tracking an Unknown Number of Targets in 2D

In this section we demonstrate the algorithm in case of an unknown number of targets moving in 2D space. The prior model for the data associations is defined such that only zero or one associations with each target on single scan is allowed.

The dynamic model for the targets is

$$\begin{pmatrix} x_{j,k} \\ y_{j,k} \\ \dot{x}_{j,k} \\ \dot{y}_{j,k} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \Delta t & 0 \\ 0 & 1 & 0 & \Delta t \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_{j,k-1} \\ y_{j,k-1} \\ \dot{x}_{j,k-1} \\ \dot{y}_{j,k-1} \end{pmatrix} + \mathbf{q}_{k-1}, \tag{4.49}$$

[1] Note that in the article (Särkkä et al., 2006b) the gamma distribution was defined such that the parameter β is here the reciprocal of the parameter in the article, i.e., $1/\beta$.

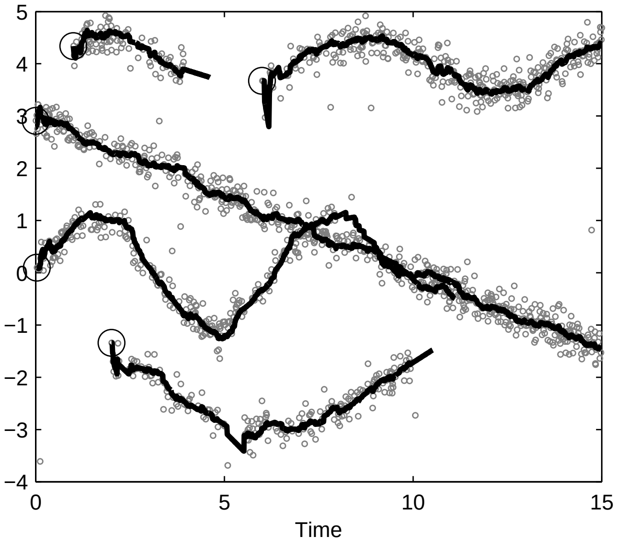

Figure 4.9: Filtering result of the 1D scenario with an unknown number of signals and parameters $\alpha = 2$ and $\beta = 1$. The circles represent the estimated starting points of the signals.

where $\mathbf{q}_{k-1}$ is process noise. The model for the measurements $\mathbf{z}_k$ is

$$\begin{aligned} z_{1,k} &= x_k + r_{x,k} \\ z_{2,k} &= y_k + r_{y,k} \end{aligned} \tag{4.50}$$

where $r_{x,k}, r_{y,k}, \sim \mathrm{N}(0, \sigma^2)$. The sampling period $\Delta t = 1/100$, process noise $q = 1/10$ in x and y directions, and measurement variance $\sigma^2 = 1/20^2$.

The detection probability of each target $p_d = 95\%$ and at each time step a random number of clutter measurements on area $[-2, 2] \times [-2.2]$ is drawn from a Poisson distribution with mean 1. The clutter measurement prior is chosen to restrict maximum the number of data associations per target to one. The birth/death model parameters are set to $p_b = 1/100$, $\alpha = 2$ and $\beta = 2$. The number of Monte Carlo samples was $N = 100$.

Figures 4.15, 4.16 and 4.17 show the filtering results, estimated number of targets and smoothed results, respectively. Again, a slight delay in the estimated target disappearance can be seen after the actual disappearance, but yet the estimated trajectories follow the true trajectories quite well. Also the estimated number of targets matches the actual number of targets well.

The restriction of maximum number of data associations per target is very important for tracking performance, especially when the amount of clutter is significant. As we have seen 100 particles is enough for in this particular scenario,

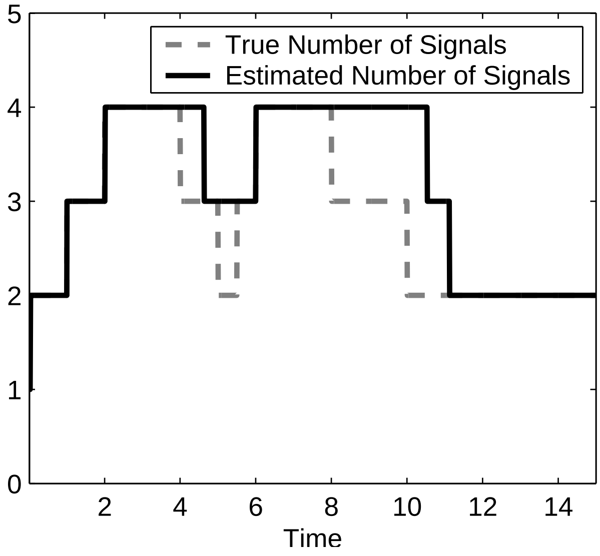

Figure 4.10: Estimated number of signals in the 1D scenario with an unknown number of signals and parameters $\alpha = 2$ and $\beta = 1$.

when the restriction is used. A quick test shows that if the restriction is not used, even with 1000 particles the result is much worse.

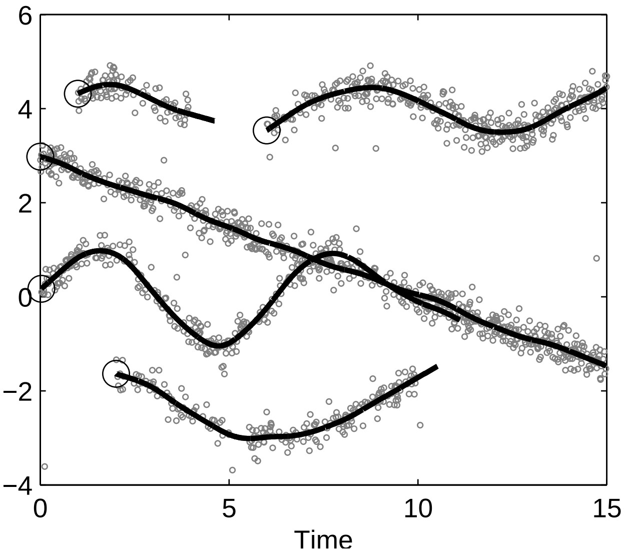

Figure 4.11: Smoothing result of the 1D scenario with an unknown number of signals and parameters $\alpha = 2$ and $\beta = 1$. The circles represent the estimated starting points of the signals.

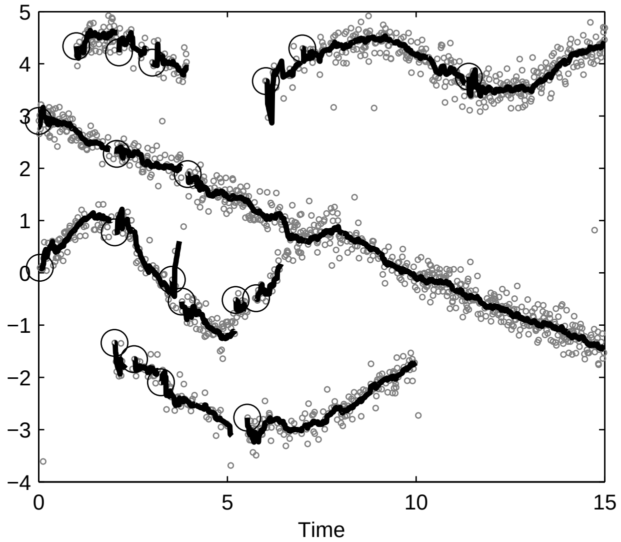

Figure 4.12: Filtering result of the 1D scenario with an unknown number of signals and parameters $\alpha = 2$ and $\beta = 10$. The circles represent the estimated starting points of the signals.

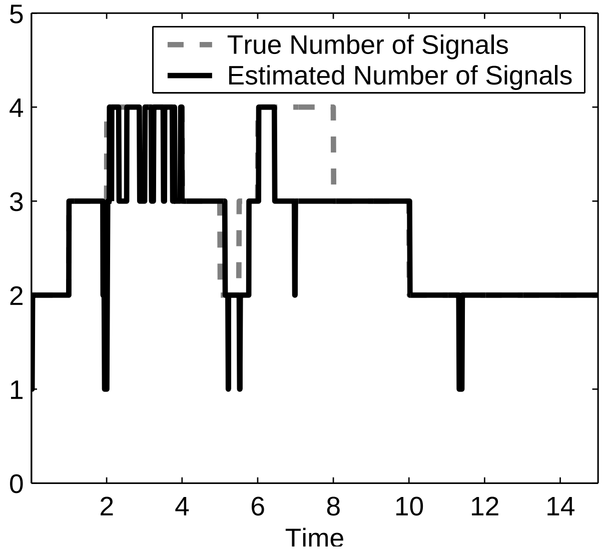

Figure 4.13: Estimated number of signals in the 1D scenario with an unknown number of signals and parameters $\alpha = 2$ and $\beta = 10$.

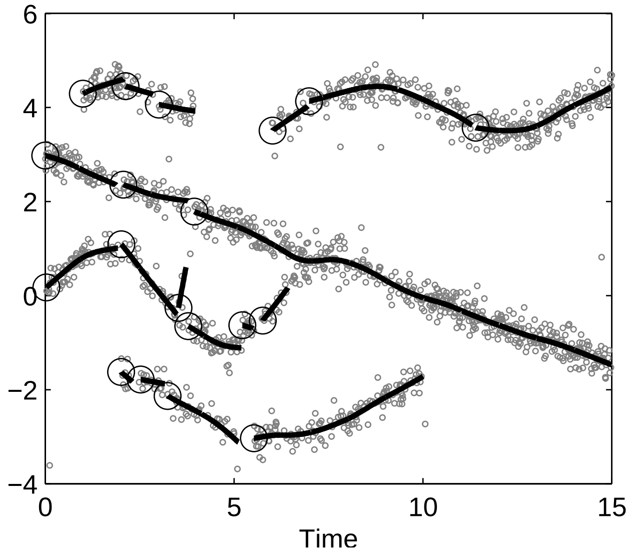

Figure 4.14: Smoothing result of the 1D scenario with an unknown number of signals and parameters $\alpha = 2$ and $\beta = 10$. The circles represent the estimated starting points of the signals.

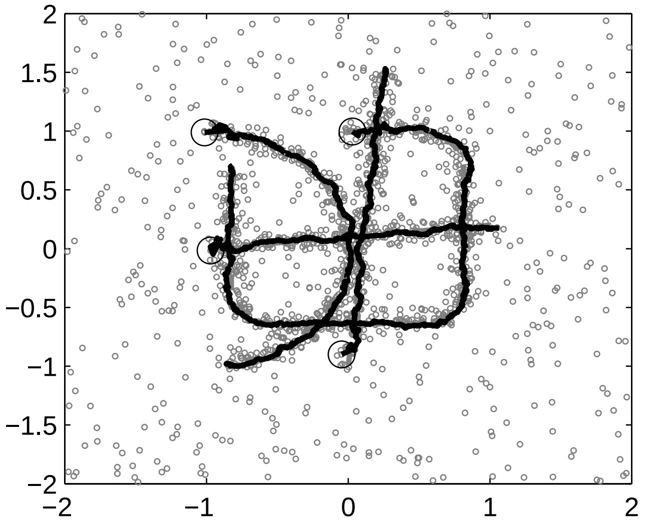

Figure 4.15: Filtering result of the 2D scenario with an unknown number of targets. The circles represent the estimated starting points of the signals.

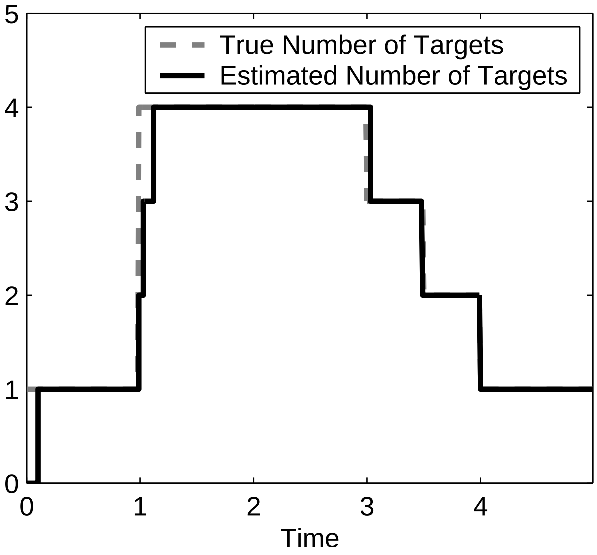

Figure 4.16: Estimated number of targets in the 2D scenario with an unknown number of targets.

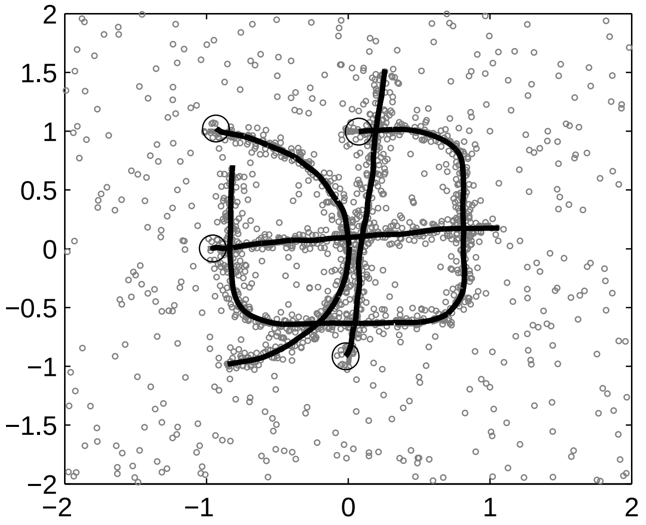

Figure 4.17: Smoothing result of the 2D scenario with an unknown number of targets. The circles represent the estimated starting points of the signals.

4.2 Spread of Infectious Diseases

This section considers the application of continuous-discrete optimal filtering and smoothing methods to estimation and prediction of spread of infectious diseases. The dynamics in this application are determined by the continuous-time model for the infection dynamics and the measurements are the recorded numbers of infected or recovered/dead individuals.

4.2.1 Classic Epidemic Model (SIR)

The classic model for the dynamics of infectious diseases is the SIR[2] model (Kermack and McKendrick, 1927; Anderson and May, 1991; Murray, 1993; Hethcote, 2000)

$$\begin{aligned} \mathrm{d}X/\mathrm{d}t &= -b\,Y\,X/N, & X(0) &= X_0, & \quad (4.51)\\ \mathrm{d}Y/\mathrm{d}t &= b\,Y\,X/N - g\,Y, & Y(0) &= Y_0, & \quad (4.52)\\ \mathrm{d}Z/\mathrm{d}t &= g\,Y, & Z(0) &= Z_0, & \quad (4.53) \end{aligned}$$

where

- $X(t)$ is the number of susceptibles at time t, that is, the number of individuals that can become infected. $X_0 \geq 0$ is the initial number of susceptibles.
- $Y(t)$ is the number of infectives, who are capable of transmitting the infection. $Y_0 \geq 0$ is the initial number of infectives.
- $Z(t)$ is the number of recovered or dead individuals, which cannot be infected anymore. $Z_0 \geq 0$ is the initial number of individuals in this class.
- $N = X(t) + Y(t) + Z(t)$ is the (constant) total number of individuals.
- b is the contact rate, which determines the rate of individuals moving from susceptible class to infectious class.
- g is the waiting time parameter such that $1/g$ is the average length of the infectious period.

Dividing the equations by the population size N yields

$$\begin{aligned} \mathrm{d}x/\mathrm{d}t &= -b\,y\,x, & x(0) &= x_0, & \quad (4.54)\\ \mathrm{d}y/\mathrm{d}t &= b\,y\,x - g\,x, & y(0) &= y_0. & \quad (4.55) \end{aligned}$$

[2] The model is called the SIR model, because the variables $X(t)$, $Y(t)$, and $Z(t)$ are often denoted as $S(t)$, $I(t)$, and $R(t)$, respectively.

where $x(t) = X(t)/N$, $y(t) = Y(t)/N$, and $z(t) = 1 - x(t) - y(t)$. Without loss of generality we shall assume $z(0) = 0$ from now on.

The analysis of the asymptotic behavior of solutions of the differential equations results in the following useful indicators, which can be used for monitoring the behavior of the model (see, e.g., Hethcote, 2000):

- The contact number $\sigma = b/g$ is the number of contacts of a typical infective during the infectious period.

- The replacement number $\sigma x(t)$ determines if there will be an epidemic or not. If $\sigma x(t) \leq 1$ then the number of infectives will decrease to zero as $t \to \infty$. If $\sigma x(t) > 1$ then the number of infectives will first increase up to a maximum and then decrease to zero.

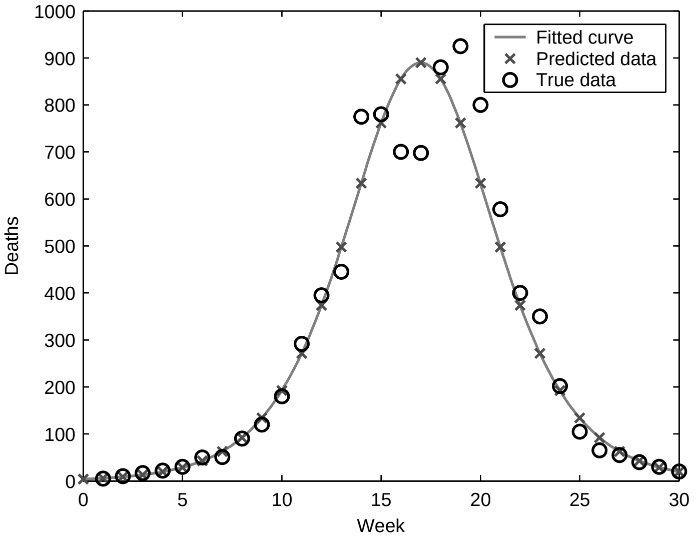

Figure 4.18: The data from Bombay plague 1905-1906 and the fit from (Kermack and McKendrick, 1927).

Example 4.1 (Bombay Plague 1905-1906)**.** *In the article (Kermack and McKendrick, 1927) the accuracy of the SIR model was demonstrated using the data from Bombay plague during the period December 1905 to July 1906. The data consists of number of deaths on each week. The model was demonstrated by fitting the parameters of an approximate solution of the differential equations to the data. The result was*

$$\frac{\mathrm{d}Z}{\mathrm{d}t} = 890\ \mathrm{sech}^2(0.2\,t - 3.4). \tag{4.56}$$

which is an approximation to the number of deaths on each week. The data and the fitted approximation are shown in the Figure 4.18.

4.2.2 Stochastic Epidemic Model

The SIR model is only an approximation to the reality and it assumes completely homogeneous mixing of individuals, that is, there are no spatial dependencies in infections. Because of this homogeneity, the dynamics of the system are determined by the time-invariant (although non-linear) differential equations. In reality there exists spatial dependencies in infection rates. In this section these unknown spatial dependencies are modeled by letting the contact number $\sigma = b/g$ depend on time in an unknown manner.

Stochastic Model of Dynamics

Because the contact number $\sigma = b/g$ is constant in the ideal model, it could be sensible to model it as a Brownian motion with a small diffusion coefficient. However, because the parameter σ is required to be positive, it is more conveniently modeled as exponential of Brownian motion. The resulting stochastic differential model is

$$\begin{aligned} \mathrm{d}x/\mathrm{d}t &= -g\,\exp(\lambda)\,y\,x \\ \mathrm{d}y/\mathrm{d}t &= g\,\exp(\lambda)\,y\,x - g\,y \\ \mathrm{d}\lambda &= q^{1/2}\,\mathrm{d}\beta, \end{aligned} \tag{4.57}$$

where $\beta(t)$ is a standard Brownian motion.

The Kolmogorov forward (or Fokker-Planck) equation corresponding to the stochastic model is

$$\frac{\partial p}{\partial t} = \frac{\partial}{\partial x}\left[g\,\exp(\lambda)\,y\,x\,p\right] + \frac{\partial}{\partial y}\left[-g\,\exp(\lambda)\,y\,x\,p + g\,y\,p\right] + \frac{1}{2}q\frac{\partial^2 p}{\partial \lambda^2}, \tag{4.58}$$

which is the equation for the probability density $p(x(t), y(t), \lambda(t)) \triangleq p(x, y, \lambda, t)$.

Prior Distribution

It is unreasonable to assume that the initial conditions $x(0)$ and $y(0)$ are known[3]. Instead, assume that the information on the initial conditions can be represented through a joint distribution $p(x(0), y(0))$. A suitable initial distribution for $x(0)$ and $y(0)$ is

$$y(0) \sim \mathrm{Beta}(\alpha_y, \beta_y), \tag{4.59}$$

$$x(0) = 1 - y(0), \tag{4.60}$$

where $\beta_y \gg \alpha_y$.

[3] The initial conditions $z(0)$ can be assumed to be zero without loss of generality.

Poisson Measurement Model

Because both in the classical SIR model and in the stochastic SIR model the values $X(t)$, $Y(t)$ and $Z(t)$ are not restricted to integer values, they cannot be interpreted as counts as such. A sensible stochastic interpretation of these values is that they are the average numbers of individuals in each class and the actual numbers of individuals are Poisson distributed with these means. Typically, either the number of newly infected or the number of recovered/dead individuals are recorded, which correspond to the following models:

- The number of newly infected individuals c_k on time period $[t_{k-1}, t_k]$ has the distribution

  $$p(c_k \,|\, \{x(\tau), y(\tau) : 0 \leq \tau \leq t_k\}, N) = \mathrm{Poisson}(c_k \,|\, N\, \phi_k), \tag{4.61}$$

  where

  $$\phi_k = x(t_{k-1}) - x(t_k). \tag{4.62}$$

- The number of the recovered/dead individuals d_k on time period $[t_{k-1}, t_k]$ has the distribution

  $$p(d_k \,|\, \{x(\tau), y(\tau) : 0 \leq \tau \leq t_k\}, N) = \mathrm{Poisson}(d_k \,|\, N\, \theta_k), \tag{4.63}$$

  where

  $$\theta_k = x(t_{k-1}) - x(t_k) + y(t_{k-1}) - y(t_k). \tag{4.64}$$

Unknown Population Size N

The model is not practical yet, because the population size N is assumed to be known. In practice, the exact value is unknown in all but simulated scenarios. However, there often exists prior information, which can be used for setting an appropriate prior distribution for the population size.

The prior information on the population size N can be modeled as a Gamma distribution

$$p(N) = \mathrm{Gamma}(N \,|\, \alpha_0, \beta_0). \tag{4.65}$$

with some suitably chosen α_0 and β_0. Note that this model does not restrict N to integer values, but it does not matter, because N is only a model parameter and there is no mathematical reason to restrict it to integer values. This particular choice of the form of prior distribution has the advantage that it allows closed form marginalization of the population size as will be seen later. Using continuous distributions as approximations to discrete distributions is common practice in Bayesian analysis (Gelman et al., 1995).

4.2.3 Estimation of the Disease

Next the implementation of the SIR filter to the stochastic SIR model is considered. The filter is constructed by the following steps:

1. A Gaussian approximation based filter (EKF) is constructed for the case of known N. This filter is later used as the importance process in SIR.
2. The equations for Rao-Blackwellization of the unknown population size are derived.
3. The equations for the scaled importance process and the likelihood ratio are derived and using these, the final SIR filter is constructed.

Gaussian Approximation

The extended Kalman filter cannot be applied to the stochastic SIR model as such, because the model is not a non-linear model driven by Gaussian noise, but instead the distribution of measurements is explicitly non-Gaussian. However, it is possible to form a Gaussian approximation to the filtering solution, which results in an algorithm that is much similar to the extended Kalman filter. The Gaussian approximation is constructed by assuming that N is known.

Assume that the state is

$$\mathbf{x}_k = \begin{pmatrix} x(t_k) \\ y(t_k) \\ \lambda(t_k). \end{pmatrix} \tag{4.66}$$

and approximately

$$p(\mathbf{x}_{k-1} \,|\, \mathbf{c}_{1:k-1}) \approx \mathrm{N}(\mathbf{x}_{k-1} \,|\, \mathbf{m}_{k-1}, \mathbf{P}_{k-1}). \tag{4.67}$$

The continuous-discrete EKF or UKF prediction can be now used for forming a Gaussian approximation to the predicted distribution

$$p(\mathbf{x}_k \,|\, \mathbf{c}_{1:k-1}) \approx \mathrm{N}(\mathbf{x}_k \,|\, \mathbf{m}_k^-, \mathbf{P}_k^-). \tag{4.68}$$

If now $\mathbf{C}$ is a matrix such that $x(t) = \mathbf{C}\,\mathbf{x}(t)$ and the distribution of measurement is

$$c_k \sim \mathrm{Poisson}(N\,x(t_{k-1}) - N\,x(t_k)), \tag{4.69}$$

then the conditional mean and variance of c_k are

$$\begin{aligned} \mathrm{E}[c_k \,|\, x] &= N\,x(t_{k-1}) - N\,x(t_k) \\ &= N\,\mathbf{C}\,\mathbf{x}_{k-1} - N\,\mathbf{C}\,\mathbf{x}_k \end{aligned} \tag{4.70}$$

$$\begin{aligned} \mathrm{Var}[c_k \,|\, x] &= N\,x(t_{k-1}) - N\,x(t_k) \\ &= N\,\mathbf{C}\,\mathbf{x}_{k-1} - N\,\mathbf{C}\,\mathbf{x}_k. \end{aligned} \tag{4.71}$$

It is now possible to form a Gaussian approximation to the posterior distribution of $\mathbf{x}_k$ as follows:

- Assume that the variance of c_k is approximately given as

$$\mathrm{Var}[c_k \,|\, x] \approx N\,\mathbf{C}\,\mathbf{m}_{k-1} - N\,\mathbf{C}\,\mathbf{m}_k^-. \tag{4.72}$$

  where $\mathbf{m}_{k-1}$ and $\mathbf{m}_k^-$ are the estimated and predicted means of the state.

- Assume that the mean is approximately given as

$$\mathrm{E}[c_k \,|\, x] \approx N\,\mathbf{C}\,\mathbf{m}_{k-1} - N\,\mathbf{C}\,\mathbf{x}_k. \tag{4.73}$$

- If we introduce an artificial measurement $c'_k = c_k - N\,\mathbf{C}\,\mathbf{m}_{k-1}$ then the measurement model can be approximated by the linear Gaussian measurement model

$$c'_k = \mathbf{H}_k\,\mathbf{x}_k + r_k, \qquad r_k \sim \mathrm{N}(0, R_k), \tag{4.74}$$

  where

$$R_k = N\,\mathbf{C}\,\mathbf{m}_{k-1} - \mathbf{C}\,\mathbf{m}_k^- \tag{4.75}$$

$$\mathbf{H}_k = -N\,\mathbf{C}. \tag{4.76}$$

The Gaussian approximation

$$p(\mathbf{x}_k \,|\, \mathbf{c}_{1:k}) \approx \mathrm{N}(\mathbf{x}_k \,|\, \mathbf{m}_k, \mathbf{P}k). \tag{4.77}$$

can be obtained by applying the Kalman filter update step using the approximate model (4.74)

The Gaussian approximation corresponding to the measurement model

$$d_k \sim \mathrm{Poisson}(d_k \,|\, N\,x(t_{k-1}) - N\,x(t_k) + N\,y(t_{k-1}) - N\,y(t_k)), \tag{4.78}$$

can be formed in analogous way except now the matrix $\mathbf{C}$ is replaced with a matrix $\mathbf{D}$ such that $x(t) + y(t) = \mathbf{D}\,\mathbf{x}(t)$.

It would be also possible to estimate the unknown N by using Gaussian approximation based filtering. The idea is to simply append the unknown N into the state and treat it as a state component having no dynamics. Gaussian approximation can be then formed in a similar manner as in the case of known N. However, this kind of joint Gaussian approximation does not work well in practice and for this reason the importance process was chosen to be formed using the Gaussian approximation where the population size was assumed to be known. An estimated value of N was used as the known value in approximation. Note that in the Gaussian approximation case it is not possible to marginalize out the parameter N, because the conjugacy of distributions disappears when Gaussian approximations are employed.

Preliminary experiments using EKF and UKF predictions showed that the EKF works better with this model. Both the EKF and UKF have stability problems, but the stability problems of UKF are more severe most likely due to the type of approximation used. The EKF probably works better, because the prediction uses the deterministic equations for mean prediction, which do preserve the normalizations and other restrictions better than the prediction of UKF.

Rao-Blackwellization of Unknown Population Size N

The model for the measurement c_1 is of the form

$$p(c_1 \mid \phi_1, N) = \mathrm{Poisson}(c_1 \mid N\, \phi_1), \tag{4.79}$$

where ϕ_1 is a deterministic function of histories of x and y.

The population size N can be marginalized out (Rao-Blackwellized) from the filtering model such that there is no need to sample it. Assume that we have just obtained the first measurement c_1. Integrating the joint distribution of c_1 and N over N gives

$$\begin{aligned} p(c_1 \mid \phi_1) &= \int p(c_1, N \mid \phi_1)\, \mathrm{d}N \\ &= \int \mathrm{Poisson}(c_1 \mid N\, \phi_1)\, \mathrm{Gamma}(N \mid \alpha_0, \beta_0)\, \mathrm{d}N \\ &= \text{Neg-bin}(c_1 \mid \alpha_0, \beta_0/\phi_1), \end{aligned} \tag{4.80}$$

which means that treating N as an unknown variable with Gamma prior turns the measurement model from Poisson to negative binomial. With given ϕ_1 the posterior distribution of N is as

$$p(N \mid c_1, \phi_1) = \mathrm{Gamma}(N \mid c_1 + \alpha_0, \phi_1 + \beta_0). \tag{4.81}$$

Thus, given the measurement c_1 and ϕ_1 the distribution of N is still a Gamma distribution. This updated distribution of N can be now used in place of the prior and the procedure can be applied recursively.

The result above can be generalized to an arbitrary step k and the more general recursions are:

- Given $p(N \mid c_{1:k-1}, \phi_{1:k-1}) = \mathrm{Gamma}(N \mid \alpha_{k-1}, \beta_{k-1})$ and the value of ϕ_k, the marginal distribution of c_k is given as

$$p(c_k \mid \phi_{1:k}) = \text{Neg-bin}(c_k \mid \alpha_{k-1}, \beta_{k-1}/\phi_k). \tag{4.82}$$

- The posterior distribution of N is then given as

$$p(N \,|\, c_{1:k}, \phi_{1:k}) = \mathrm{Gamma}(N \,|\, c_k + \alpha_{k-1}, \phi_k + \beta_{k-1}). \tag{4.83}$$

The recursion above can be used for forming a Rao-Blackwellized particle filter for the stochastic model. The result applies to the measurements d_k in an analogous manner:

- Given $p(N \,|\, d_{1:k-1}, \theta_{1:k-1}) = \mathrm{Gamma}(N \,|\, \alpha_{k-1}, \beta_{k-1})$ and the value of θ_k the marginal distribution of d_k is given as

$$p(d_k \,|\, \phi_{1:k}) = \text{Neg-bin}(d_k \,|\, \alpha_{k-1}, \beta_{k-1}/\theta_k). \tag{4.84}$$

- The posterior distribution of N is then given as

$$p(N \,|\, d_{1:k}, \phi_{1:k}) = \mathrm{Gamma}(N \,|\, d_k + \alpha_{k-1}, \theta_k + \beta_{k-1}). \tag{4.85}$$

Continuous-Discrete SIR Implementation

The approximate Gaussian solution can be used as the importance process in a continuous-discrete sequential importance sampling filter. Because the population size can be easily estimated from the current values of α_k and β_k the Gaussian approximation can be formed by assuming that N is known.

Assume that the state vector is defined as

$$\mathbf{x}_k = \begin{pmatrix} x(t_k) \\ y(t_k) \\ \lambda(t_k). \end{pmatrix} \tag{4.86}$$

When $\mathbf{x}_{k-1} = \mathbf{x}(t_{k-1})$ is given, by using the extended Kalman filter like algorithm it is now possible to form a Gaussian approximation to the distribution of $\mathbf{x}(t_k)$. If the approximate marginal distribution of the variable $\lambda(t)$ is $\mathrm{N}(m_{3,k}, P_{3,k})$ then it is possible to form an importance process for it as

$$\mathrm{d}s = \frac{m_{3,k} - x_{3,k}}{\Delta t}\,\mathrm{d}t + \sqrt{\frac{P_{3,k}}{\Delta t}}\,\mathrm{d}\beta, \tag{4.87}$$

where $\Delta t = t_k - t_{k-1}$. The Algorithm 3.29 can be now used for performing the sequential importance sampling. Because the stochastic equation of $\lambda(t)$ has no drift term, the likelihood ratio has now the following simple form:

$$Z(t;\omega) = \exp\left(\frac{(m_{3,k} - x_{3,k})}{\sqrt{P_{3,k}\,\Delta t}} (\beta(t_{k-1}) - \beta(t_k)) - \frac{1}{2}\frac{(m_{3,k} - x_{3,k})^2}{P_{3,k}} \right) \tag{4.88}$$

The continuous-discrete SIR filter can be now implemented as follows. Given a set of weighted particles

$$\{w_{k-1}^{(i)}, x_{k-1}^{(i)}, y_{k-1}^{(i)}, \lambda_{k-1}^{(i)}, \alpha_{k-1}^{(i)}, \beta_{k-1}^{(i)} \,:\, i = 1, \ldots, n\}, \tag{4.89}$$

representing the distribution of population at time step t_{k-1} the SIR filter processes the measurement c_k (or d_k) as follows:

1. For each particle, use the Gaussian approximation based filter for forming Gaussian approximation to the joint posterior distribution of state variables $(x^{(i)}(t_k), y^{(i)}(t_k), \lambda^{(i)}(t_k))^T$ given the measurement.

2. Form a predicted particle set by simulating the scaled importance process (4.87) from t_{k-1} to t_k using the Gaussian approximations above and using each particle as a starting point. This results in the particles $\{x_k^{(i)}, y_k^{(i)}, \lambda_k^{(i)}\}$.

3. Compute the likelihood ratio by Equation (4.88) using the same simulated Brownian motions as in simulation of the importance process above. This results in likelihood ratios $Z_k^{(i)}$.

4. Compute weights as follows (if measurement is c_k):

$$w_k^{(i)} \propto w_{k-1}^{(i)}\, Z_k^{(i)}\, \text{Neg-bin}(c_k \,|\, \alpha_{k-1}^{(i)}, \beta_{k-1}^{(i)} / \phi_k^{(i)}), \tag{4.90}$$

   where $\phi_k^{(i)} = x_k^{(i)} - x_{k-1}^{(i)}$ or as follows (if measurement is d_k):

$$w_k^{(i)} \propto w_{k-1}^{(i)}\, Z_k^{(i)}\, \text{Neg-bin}(d_k \,|\, \alpha_{k-1}^{(i)}, \beta_{k-1}^{(i)} / \theta_k^{(i)}), \tag{4.91}$$

   where $\theta_k^{(i)} = x^{(i)}(t_{k-1}) - x^{(i)}(t_k) + y^{(i)}(t_{k-1}) - y^{(i)}(t_k)$.

5. Update the population size parameters for all i (if measurement is c_k):

$$\begin{aligned} \alpha_k^{(i)} &= \alpha_{k-1}^{(i)} + c_k \\ \beta_k^{(i)} &= \beta_{k-1}^{(i)} + \phi_k^{(i)}. \end{aligned} \tag{4.92}$$

   or (if measurement is d_k):

$$\begin{aligned} \alpha_k^{(i)} &= \alpha_{k-1}^{(i)} + d_k \\ \beta_k^{(i)} &= \beta_{k-1}^{(i)} + \theta_k^{(i)}. \end{aligned} \tag{4.93}$$

6. Resample by deleting/duplicating particles proportionally to the weights.

4.2.4 Results

Bombay Plague

The continuous-discrete SIR filter described in this section was applied to the classical Bombay plague data shown in Figure 4.18. The EKF based Gaussian approximation was used as the importance process and 10000 particles was used. The prior distribution for proportion of initial infectives was Beta$(1, 100)$. The population size prior was Gamma$(10, 0.001)$. The waiting time parameter was assumed to be $g = 1$. The prior distribution for $\lambda(0)$ was $\mathrm{N}(\ln(5), 4)$. The diffusion coefficient of the Brownian motion was $q = 0.001$. Between measurements the EKF predictions were integrated using 5 steps of fourth order Runge-Kutta integration and the importance processes and likelihood ratios were simulated using the stochastic Runge-Kutta method.

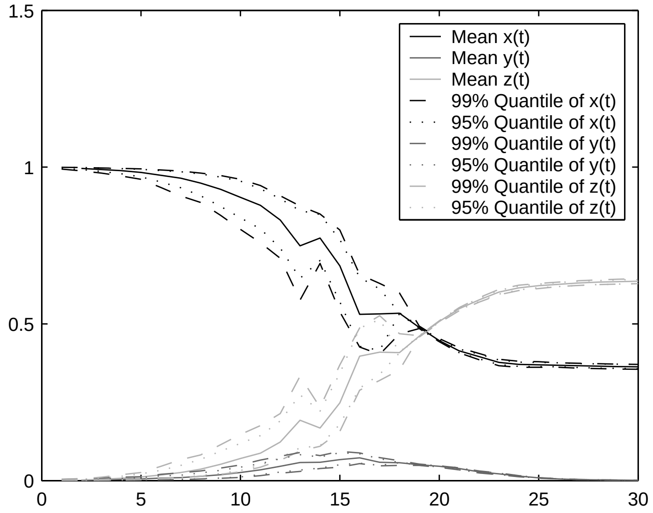

Figure 4.19: Filtered estimates of values of $x(t)$, $y(t)$, and $z(t)$ from the Bombay data.

The final filtered estimates of the histories of $x(t)$, $y(t)$, and $z(t)$ are shown in Figure 4.19. These estimates are filtered estimates, that is, they are conditional to the previously observed measurements only. That is, the estimate on week t is the estimate that could be actually computed on week t without any knowledge of the future observations. The estimates look quite much as what would be expected. The proportion of susceptibles decreases monotonically, the number of infectives increases up to a maximum and then decreases to zero. However, these estimated values are not very useful themselves. The reason for this is that, for example, the value x_∞ which is the remaining value of susceptibles in the end depends on the choice of g and other prior parameters. That is, these estimated values are not absolute in the sense that their values depend heavily on the prior assumptions.

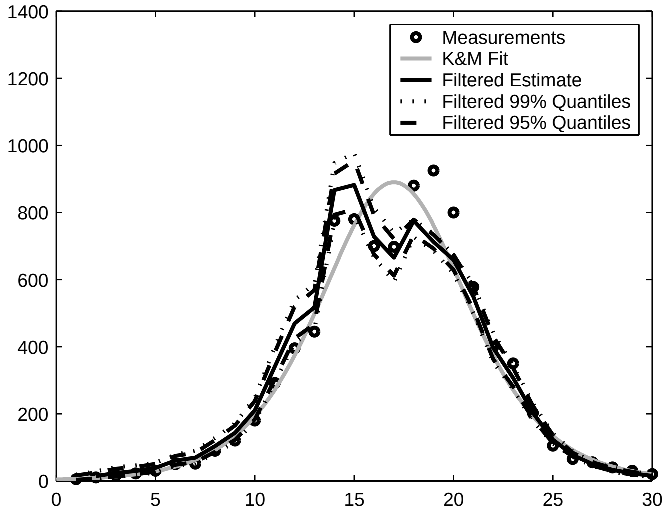

Figure 4.20: Filtered estimate of $\mathrm{d}Z/\mathrm{d}t$ from the Bombay data. The estimate of (Kermack and McKendrick, 1927) is also shown for comparison.

Much more informative quantity is the value $\mathrm{d}Z/\mathrm{d}t$, whose filtered estimate is shown in Figure 4.20. The classical estimate presented in (Kermack and McKendrick, 1927) is also shown. The SIR filter estimate can be seen to differ a bit from the classical estimate, but still both the estimates look quite much like what would be expected. Note that the classical estimate is based on all measurements, whereas the filtered estimate is based on observations made up to that time only. That is, the filter estimate could be actually computed on week t, but the classical estimate could not.

The filtered estimates of values $\sigma(t)$ are shown in Figure 4.21. The value can be seen to vary a bit on time, but the estimated expected value remains on the range $[1.4, 1.8]$ all the time. As can be seen from the figure, according to the data the value of $\sigma(t)$ is not constant. This is not surprising, because the spatial and other unknown effects are not accounted at all in the classical SIR model and these effects typically affect the number of contacts.

A very useful indicator value is $\sigma(t)\,x(t)$, whose filtered estimate is shown in Figure 4.22. In the deterministic SIR model with constant σ this indicator defines the asymptotic behavior of the epidemic (see, e.g., Hethcote, 2000): If $\sigma x(t) \leq 1$ then the number of infectives will decrease to zero as $t \to \infty$. If $\sigma x(t) > 1$ then the number of infectives will first increase up to a maximum and then decrease to zero. As can be seen from the Figure 4.22 the filtered estimate of the indicator

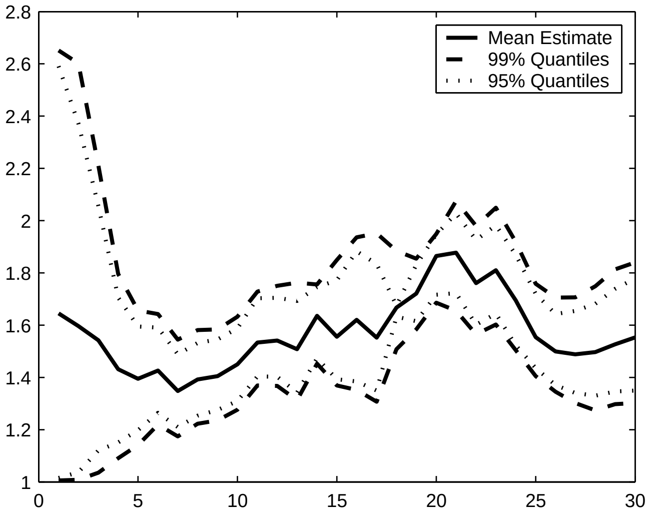

Figure 4.21: Bombay plague: Filtered estimate of value $\sigma(t)$.

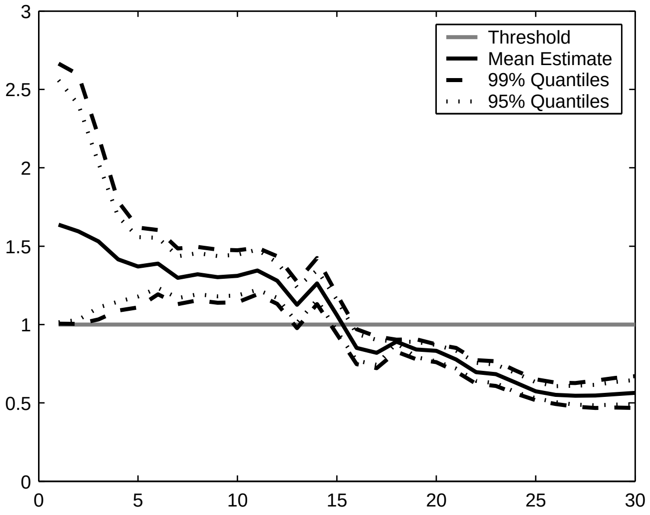

Figure 4.22: Bombay plague: Filtered estimate of value $x(t)\,\sigma(t)$.

value goes below 1 just after the maximum somewhere between weeks 15–16, which can be seen in Figure 4.20. That is, the estimated value of $\sigma(t)\,x(t)$ could be used as an indicator, which tells if the epidemic is over or not.

Using the particles it is also possible to predict ahead to the future and estimate

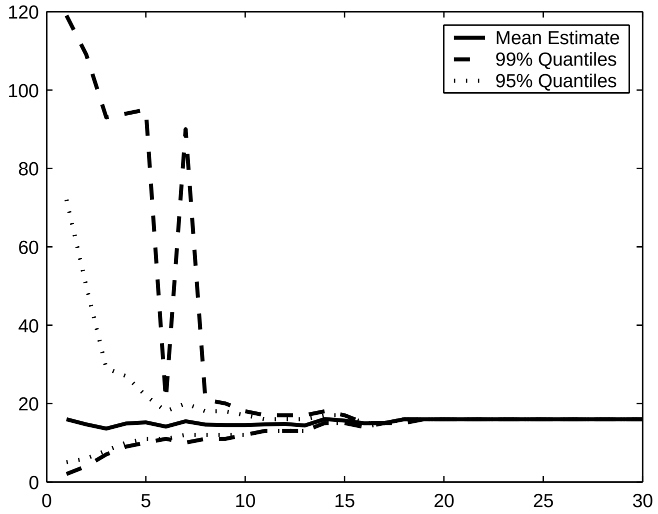

Figure 4.23: Bombay plague: Filtered estimate of time of maximum of epidemic.

the time when the maximum of the epidemic will be reached. The estimate computed from the filtering result is shown in the Figure 4.23. Again, the estimates are filtered estimates and the estimate on week t could be actually computed on week t, because it depends only on the counts observed up to that time. The filtered estimate can be seen to quickly converge to the values near the correct maximum on weeks 15–16. It is interesting to see that the prediction is quite accurate already around the week 10, which is far before reaching the actual maximum. If this kind of prediction had been done on, for example, week 10 of the disease, it would have predicted the time of actual epidemic maximum quite accurately. After the maximum has been observed, the estimate quickly converges to a constant value, which according to the Figure 4.20 is likely to be near the true maximum.

A very useful estimate is also the expected total number of deaths caused by the epidemic. This can be computed from the filtered estimates and the result is shown in Figure 4.24. In the beginning the estimate is very diffuse, but after maximum has been reached the estimate converges near the correct value. The estimate is a bit less than the observed value long before reaching the maximum, which might be due to existence of two maximums in the observed data (see, Figure 4.20). Because the second maximum is not predicted by the model, the extra number of deaths caused by it cannot be seen in the predictions.

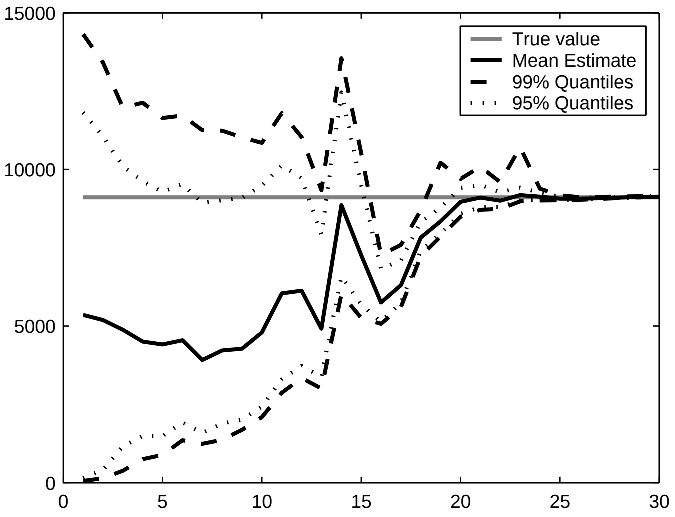

Figure 4.24: Bombay plague: Filtered estimated of number of deaths.

4.3 CATS Time Series Prediction Competition

This section presents the winning solution to the time series prediction competition, the CATS benchmark (Lendasse et al., 2004), which was organized as a special session of the IJCNN 2004 conference. The original solution was presented in article (Särkkä et al., 2004b) and extended results in (Särkkä et al., 2006a). The solution is based on the classical Kalman smoother with cross-validated process noise variances.

4.3.1 CATS Benchmark

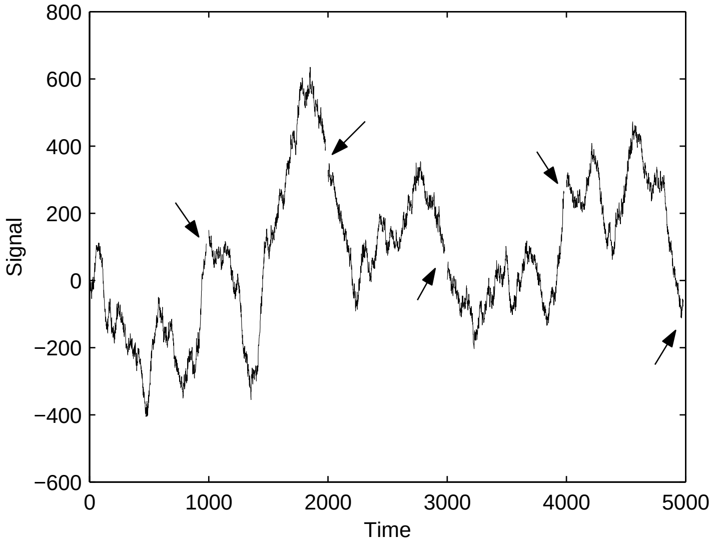

Figure 4.25: The CATS benchmark time series. The purpose of the competition was to predict the missing data (marked with arrows) such that the mean squared error is minimized.

The goal of the CATS competition (Lendasse et al., 2004) was to provide a new benchmark for the problem of time series prediction and to compare different methods and models that can be used for the prediction. The proposed time series is the *CATS benchmark* (Competition on Artificial Time Series).

This artificial time series with 5,000 data was given. Within those 100 values were missing. These missing values were divided in 5 blocks:

- elements 981 to 1,000;
- elements 1,981 to 2,000;

- elements 2,981 to 3,000;
- elements 3,981 to 4,000;
- elements 4,981 to 5,000;

The purpose was to predict the 100 missing values based on the other data. The performance criterion was the mean square error, which was computed on the 100 missing values. The time series is shown in Figure 4.25.

Relationship of Gaussian Process Regression and Kalman Filtering

The continuous-discrete Kalman filter uses *Gaussian processes* as its dynamic models. Gaussian processes are the most common signal models in classical continuous-time signal processing, especially in communications applications. The theory of Wiener filtering (Wiener, 1950), which is the theoretical basis for optimal signal detection and demodulation (Van Trees, 1968) deals with signals that can be modeled as stationary Gaussian processes. Kalman filtering (Kalman, 1960b) and Kalman-Bucy filtering (Kalman and Bucy, 1961) can be considered as extensions to Wiener filtering theory, in which also non-stationary Gaussian process models can be used. Stochastic control theory (Maybeck, 1982b) builds on the grounds of Kalman-Bucy filtering by including a controller aside with the optimal state estimator.

Gaussian processes, or *Gaussian random fields* are also used in spatial and spatio-temporal modeling (Christakos, 1992; Banerjee et al., 2004), and in general regression and classification problems (O'Hagan, 1978; Williams and Rasmussen, 1996; Barber and Williams, 1997; MacKay, 1998; Neal, 1999). Furthermore, the functional prior implied by an MLP neural network model converges to a Gaussian process as the number of hidden units increases, provided that the MLP weight priors are chosen suitably (Neal, 1996).

The relationship between the Gaussian processes used in regression and the Gaussian processes used in filtering is that continuous-discrete filtering can be thought of as regression from time t to partially observed states $\mathbf{x}(t)$, which we observe through the measurements $\mathbf{y}(t)$. The Gaussian process dynamic model is the prior for the functions $t \mapsto \mathbf{x}(t)$. The optimal filter solves the state estimates recursively at each time instance, and it can be considered the on-line learning solution to the Gaussian process regression problem. However, the filter provides the on-line estimates only forward in time, not at arbitrary time instances, and to compute the state estimates at arbitrary time instances the smoothing step is required. The Gaussian processes used in regression have multidimensional time-variables (i.e., regressors) and for this reason they are often called Gaussian random fields.

4.3.2 Description of the Model

The Long Term Model

For long term prediction, a linear dynamic model is likely to be a good approximate model because if we ignore the short term periodicity of the data, the data could be well generated by a locally linear Gaussian process with Gaussian measurement noise. The data seems to consist of lines with suddenly changing derivatives. Thus, it would be reasonable to model the derivative as Brownian noise process, which leads to a white noise model for the second derivative. Using higher derivatives does not seem useful, because the curve consists of a set of straight lines rather than parabolas or other higher order curves.

The dynamic model is formulated as a continuous time model, and then discretized to allow for a varying sampling rate, that is, prediction over the missing measurements. The selected dynamic linear model for the long term prediction is the stochastic differential equation model

$$\frac{d^2 x(t)}{dt^2} = w(t), \tag{4.94}$$

where $w(t)$ is a continuous-time Gaussian white noise process with moments

$$\begin{aligned} \mathrm{E}[w(t)] &= 0 \\ \mathrm{E}[w(t)\, w(t+\tau)] &= q^x \delta(\tau). \end{aligned} \tag{4.95}$$

This can be written in equivalent discrete form as

$$\begin{pmatrix} x_k \\ \dot{x}_k \end{pmatrix} = \begin{pmatrix} 1 & \Delta t \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_{k-1} \\ \dot{x}_{k-1} \end{pmatrix} + \begin{pmatrix} q^x_{1,k-1} \\ q^x_{2,k-1} \end{pmatrix}, \tag{4.96}$$

where the process noise, $\mathbf{q}^x_k = (q^x_{1,k-1} \; q^x_{2,k-1})^T$, has zero mean and covariance

$$\mathbf{Q}_{k-1} = \begin{pmatrix} \Delta t^3/3 & \Delta t^2/2 \\ \Delta t^2/2 & \Delta t \end{pmatrix} q^x, \tag{4.97}$$

and where Δt is the time period between samples and q^x defines the strength (spectral density) of the process noise. The measurement model is

$$y_k = x_k + r^x_k, \quad r^x_k \sim \mathrm{N}(0, \sigma^2_x). \tag{4.98}$$

A quick testing of the long term model produces a smooth curve as shown in Figure 4.26. It can be seen that the locally linear dynamic model may be a bit too simple, because the residual signal still seems to contain noticeable periodicity. This periodicity can be best seen from the residual autocorrelation in Figure 4.27.

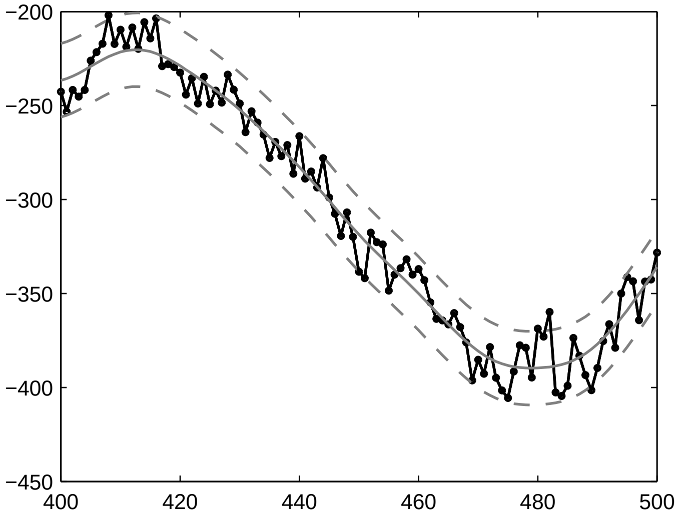

Figure 4.26: Data 400–500 (black) and the result of prediction with the long term model (gray).

The Short Term Model

The short term periodicity of the residual time series $\{e_k : k = 1, ..., N\}$ can be modeled with a time varying autoregressive (TVAR) model (West and Harrison, 1997), in which as an extension to conventional AR models (Hayes, 1996), the weights are allowed to vary according to a Gaussian random walk model

$$\begin{aligned} \mathbf{w}_k &= \mathbf{w}_{k-1} + \mathbf{v}_k^{\mathsf{ar}} \\ e_k &= \sum_i w_{i,k} e_{k-i} + r_k^{\mathsf{ar}}. \end{aligned} \tag{4.99}$$

The process noise $\mathbf{v}_k^{\mathsf{ar}}$ has zero mean and covariance $\mathbf{Q} = q^{\mathsf{ar}}\mathbf{I}$. The weight vector $\mathbf{w}_k$ is estimated from the known part of the residual time series. The measurement noise has a Gaussian distribution $r_k^{\mathsf{ar}} \sim \mathrm{N}(0, \sigma_{\mathsf{ar}}^2)$. A second order AR-model was chosen such that the weight vector was two dimensional,

$$\mathbf{w}_k = \begin{pmatrix} \mathbf{w}_{1,k} \\ \mathbf{w}_{2,k} \end{pmatrix}. \tag{4.100}$$

After the TVAR-model weights have been estimated from the residual time series data, the final estimation solution is obtained by estimating the short term

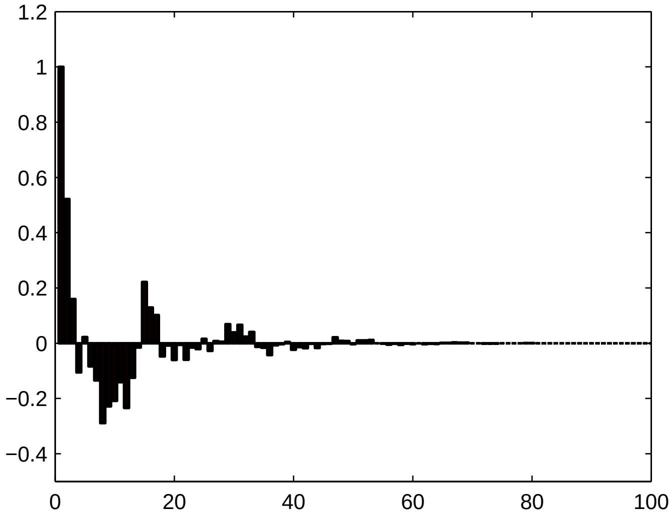

Figure 4.27: Autocorrelation in the residual of the long term prediction model.

signal d_k from the model

$$\begin{aligned} d_k &= \sum_i w_{i,k} d_{k-i} + v_k^{\mathsf{p}} \\ e_k &= d_k + r_k^p, \quad r_k^p \sim \mathrm{N}(0, \sigma_p^2), \end{aligned} \tag{4.101}$$

where the process noise v_k^{p} has variance q^{p}. The final signal estimate is then given as $\hat{y}_k = \hat{x}_k + \hat{d}_k$, where $\hat{x}_k$ is the estimate produced by applying Kalman smoother to the long term model, and $\hat{d}_k$ is produced by the short term model.

In practice only the distributions of weight vectors $\mathbf{w}_k$ are known, not their actual values, and in order to use the model (4.101) we would have to integrate over these distributions at every time step. This integration was approximated by using the most likely estimate of the weight vector time series with this single estimate regarded as being known in advance. In classical statistical signal processing this estimate is calculated by linear least squares (see, e.g., Hayes, 1996). Because here the weight vector is allowed to vary in time, in this case the corresponding estimate is produced by applying the Kalman smoother to the model (4.99).

The Prediction Method

The long term prediction is done in two steps:

1. Run the *Kalman filter* over the data sequence and store the estimated means

and covariances. Predict the missing measurements such that the filtering result contains estimates also for the missing steps.

2. Run the *Kalman smoother* over the Kalman filter estimation result, which results in the smoothed (MAP) estimate of the time series including the missing parts.

The short term prediction consists of four steps:

1. Run the *Kalman filter* over the residual sequence with the model (4.99) in order to produce a filtering estimate of the TVAR weight vectors. Predict the weights over the missing parts.

2. Run the *Kalman smoother* over the Kalman filter estimation result above, which results in a smoothed (MAP) estimate of the weight time series including the missing parts.

3. Run the *Kalman filter* over the residual sequence with the model (4.101) in order to produce a filtering estimate of the short term periodicity. The periodicity is also predicted over the missing parts.

4. Run the *Kalman smoother* over the Kalman filter estimation result above, which results in a smoothed (MAP) estimate of the periodicity time series including the missing parts.

Due to the Gaussian random walk model of the weights the short term model potentially has a large effective number of parameters. A simple error minimization procedure with respect to the noise parameters (e.g., Maximum Likelihood) would lead to a badly over-fitted estimation solution. By applying cross-validation the predictive performance can be maximized and the overfilling can be avoided.

4.3.3 The Original Results

Selection of Measurement Noises

The long term measurement noise strength can be approximated by looking at a short time period of the curve. Assuming that it was approximated it with a dynamic linear model, we could approximate the standard deviation of the model's measurement noise by looking at the strengths of the residuals. The selected variance of the noise was $\sigma_x^2 = 10^2$, which fits to the observed residual as can be seen in the Figure 4.26 quite well.

The choices of the measurement noises both in the long and the short term models can be done, for example, by visual inspection, because the exact choice of the noise strengths is not crucial. In fact, the choice does not matter at all when the cost function of the CATS competition is considered, because in this case the selection of measurement noise strength is dependent on the selection of the process noise strength in all the models. The process noise strength is selected based on cross-validation, which implicitly corrects also the choice of the measurement noise strength. By visual inspection the suitable measurement noise for the TVAR-estimation model (4.99) was $\sigma^2_{\mathsf{ar}} = 1^2$.

Because we are only interested in the missing parts of data in prediction with the model (4.101), the best way to do this is to follow the measurements exactly whenever there are measurements and use the TVAR-model for prediction only when there are no measurements. This happens when the measurement noise level is set to as low as possible and the process noise is set to a moderate value. The choice for the measurement noise level in model (4.101) was $\sigma^2_{\mathsf{p}} = 10^{-9}$.

Cross-Validation of The Process Noises

The process noise parameters q^x and q^{ar} were selected using a decision theoretic approach by minimizing the expected cost, where the cost function was the target error criterion. The expected cost can easily be computed by cross-validation, which approximates the formal Bayes procedure of computing the expected costs.

Cross-validation methods for model selection have been proposed by several authors: for early accounts see (Stone, 1974; Geisser, 1975) and for a more recent review see (Gelfand et al., 1992; Shao, 1993). Bernardo and Smith (1994) and Vehtari and Lampinen (2002) discuss how cross-validation approximates the formal Bayes procedure of computing the expected utility of using a model for predictions.

Based on the cross-validation, the best process noises were

$$\begin{aligned} q^x &= 0.14 \\ q^{\mathsf{ar}} &= 0.0005. \end{aligned} \tag{4.102}$$

These values were based on cross-validation over a range of values, which was selected in advance. However, it later turned out that this range could have been selected better. As already discussed in this section, the only requirement for the selection of the process noise q^{p} is that it should be high enough. Because the measurement noise was chosen to be very low, our choice was $q^{\mathsf{p}} = 1$.

The Original Prediction Results

Figure 4.28 shows the estimated TVAR-coefficients for each time instance. It can be seen that the weights vary a bit over time, but the periodic short term process seems to be quite stationary.

Figures 4.29, 4.30, 4.31, 4.32 and 4.33 show the results of predicting over the missing intervals. It can be seen that on the missing intervals the short term model differs from the long term model only near the measurements and the combined

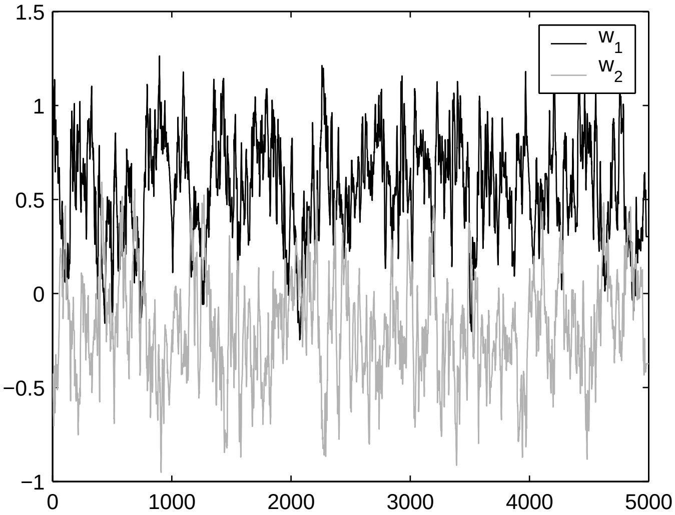

Figure 4.28: Estimated filter coefficients for the TVAR-model.

estimate is closest to the long term prediction in the middle of the prediction period. The result is intuitively sensible, because when we are going away from the measurements, we have less information about the phase of the local periodicity, and it is best just to guess the mean given by the long term model.

The following mean squared errors were obtained by the method:

$$\begin{aligned}
E_1 = \frac{1}{100}\Bigg[&\sum_{t=981}^{1000}(y_t-\hat{y}_t)^2 + \sum_{t=1981}^{2000}(y_t-\hat{y}_t)^2 + \sum_{t=2981}^{3000}(y_t-\hat{y}_t)^2 \\
&+ \sum_{t=3981}^{4000}(y_t-\hat{y}_t)^2 + \sum_{t=4981}^{5000}(y_t-\hat{y}_t)^2\Bigg] = 408 \\
E_2 = \frac{1}{80}\Bigg[&\sum_{t=981}^{1000}(y_t-\hat{y}_t)^2 + \sum_{t=1981}^{2000}(y_t-\hat{y}_t)^2 \\
&+ \sum_{t=2981}^{3000}(y_t-\hat{y}_t)^2 + \sum_{t=3981}^{4000}(y_t-\hat{y}_t)^2\Bigg] = 346.
\end{aligned} \tag{4.103}$$

The error E_1 was the actual CATS competition objective, and the approach described in this section gave the lowest error in the competition (Lendasse et al., 2004). The second error criterion was used in further analysis on the different methods in (Lendasse et al., 2004).

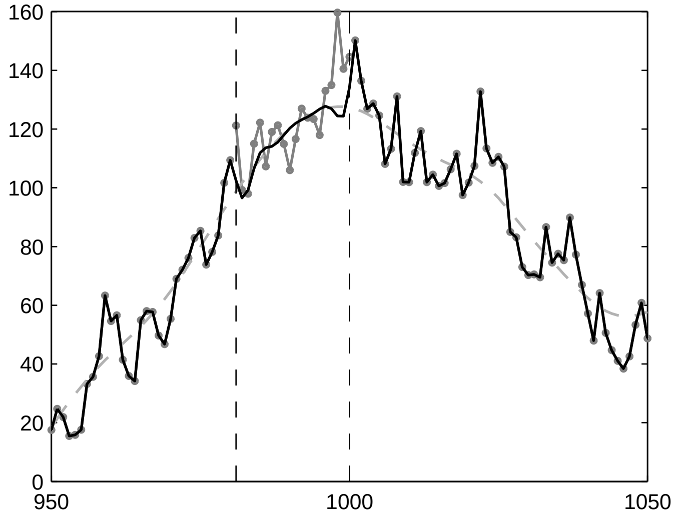

Figure 4.29: The original prediction over missing data at 981 – 1000. The gray line is the true signal, the dashed line is the long term prediction result, and the black line is the combined long and short term prediction result.

4.3.4 Improved Results

Extended Cross-Validation of Process Noises

Further analysis of the cross-validation results of the original prediction competition indicated that if we had used a larger range of possible noise levels in the cross-validation, we would have obtained a better prediction result. Extending the cross-validation to zero noise levels reveals that based on the cross-validation the following parameters are better than the original ones used in the competition:

$$\begin{aligned} q^x &= 0.14 \\ q^{\mathsf{ar}} &= 0. \end{aligned} \tag{4.104}$$

The difference to the original parameters is that the TVAR model process noise level is exactly zero, not only almost zero as in the original model. This means that it is better to use a stationary AR-model, not a time-varying AR model in prediction. Fitting the AR model to the data with the cross-validated noise parameters resulted in the following AR-parameters:

$$\begin{aligned} w_1 &= 0.6089 \\ w_2 &= -0.1517. \end{aligned} \tag{4.105}$$

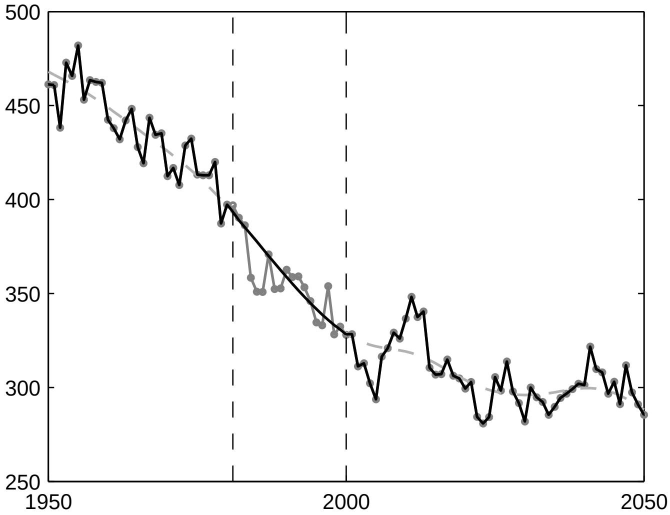

Figure 4.30: The original prediction over missing data at 1981 – 2000. The gray line is the true signal, the dashed line is the long term prediction result, and the black line is the combined long and short term prediction result.

The error criteria are also better than in the original predi ction:

$$\begin{aligned} E_1 = \frac{1}{100}\Bigg[&\sum_{t=981}^{1000}(y_t-\hat{y}_t)^2+\sum_{t=1981}^{2000}(y_t-\hat{y}_t)^2+\sum_{t=2981}^{3000}(y_t-\hat{y}_t)^2 \\ &+\sum_{t=3981}^{4000}(y_t-\hat{y}_t)^2+\sum_{t=4981}^{5000}(y_t-\hat{y}_t)^2\Bigg]=381 \\ E_2 = \frac{1}{80}\Bigg[&\sum_{t=981}^{1000}(y_t-\hat{y}_t)^2+\sum_{t=1981}^{2000}(y_t-\hat{y}_t)^2 \\ &+\sum_{t=2981}^{3000}(y_t-\hat{y}_t)^2+\sum_{t=3981}^{4000}(y_t-\hat{y}_t)^2\Bigg]=312. \end{aligned} \tag{4.106}$$

The Improved Prediction Results

Figures 4.34, 4.35, 4.36, 4.37 and 4.38 show the results of predicting over the missing intervals with the improved method. The difference in the result is that in the improved prediction the AR model seems to get slightly less weight and the long term prediction dominates more.

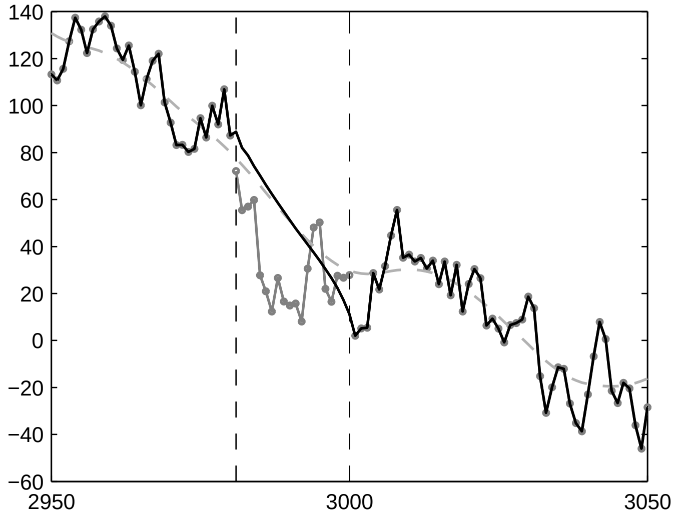

Figure 4.31: The original prediction over missing data at 2981 – 3000. The gray line is the true signal, the dashed line is the long term prediction result, and the black line is the combined long and short term prediction result.

4.3.5 Summary of Prediction Results

In this section the winning solution to the CATS time series prediction competition has been presented. The solution is based on applying the classical Kalman smoother method to estimating the long term and short term statistical models for the CATS benchmark time series. The good prediction performance is likely due to that the long term prediction gives a very good overall approximation of the signal and the short term prediction catches the local periodicity ignored by the long term model.

It has also been shown that the original prediction results can be improved by simplifying the model, namely by removing the time-dependence from the AR model. This model choice would have turned out also in the original cross-validation if a bit larger parameter range in the cross-validation had been used.

Although all the used models were linear (and dynamic) in nature they seem to model this non-linear time series well. The good performance is not surprising, because the long term Gaussian process model is very much related to the Gaussian processes which have obtained good results in Bayesian non-linear regression. The short term prediction model is also a Gaussian process model, but of the type which is more common in the signal processing context.

It could be possible that by using some kind of non-linear state space models (filtering models) the prediction results would be better, but it is very hard to

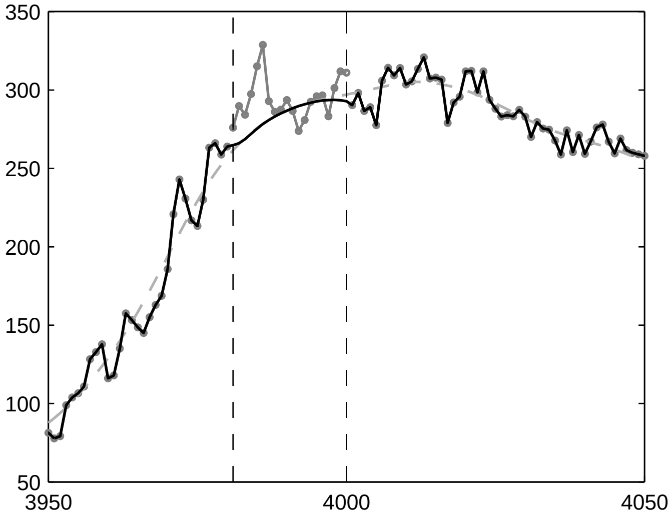

Figure 4.32: The original prediction over missing data at 3981 – 4000. The gray line is the true signal, the dashed line is the long term prediction result, and the black line is the combined long and short term prediction result.

judge what kind of model really is the best. Using more complex models would restrict the generality of the approach and even though some specific models could improve the results with this particular time series, finding generally better models than the classical Gaussian process models is difficult.

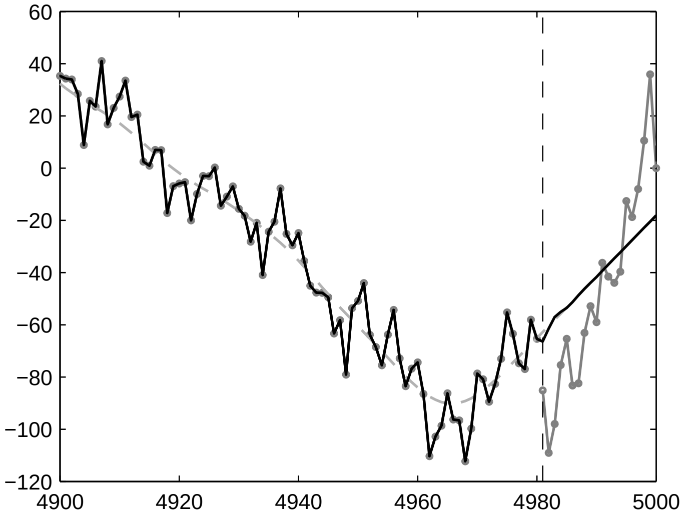

Figure 4.33: The original prediction over missing data at 4981 – 5000. The gray line is the true signal, the dashed line is the long term prediction result, and the black line is the combined long and short term prediction result.

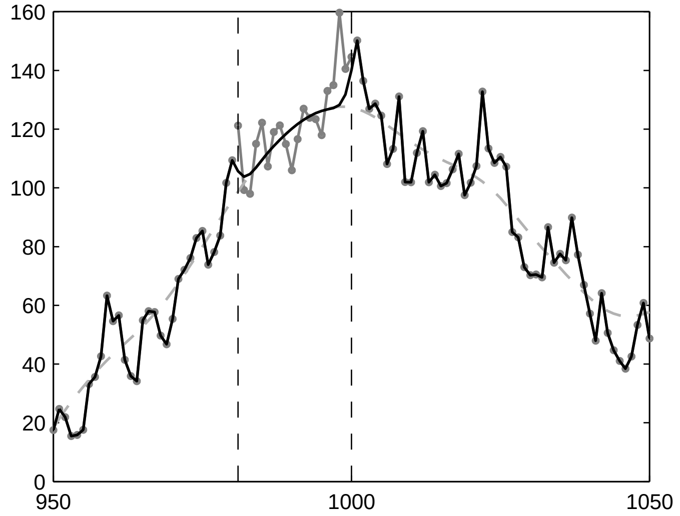

Figure 4.34: The improved prediction over missing data at 981 – 1000.

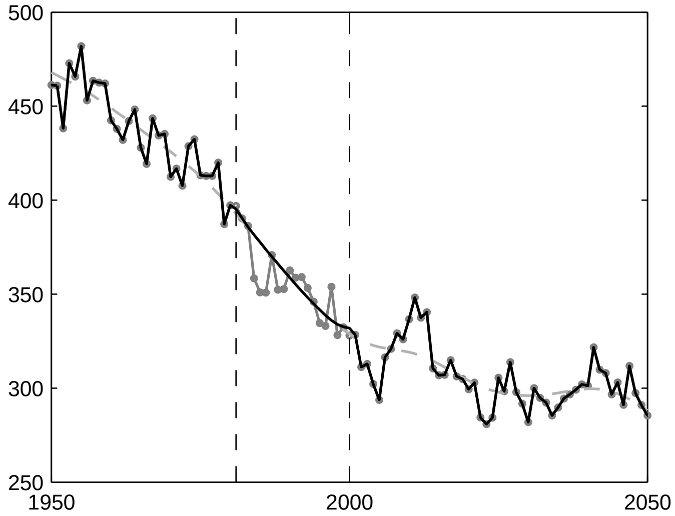

Figure 4.35: The improved prediction over missing data at 1981 – 2000.

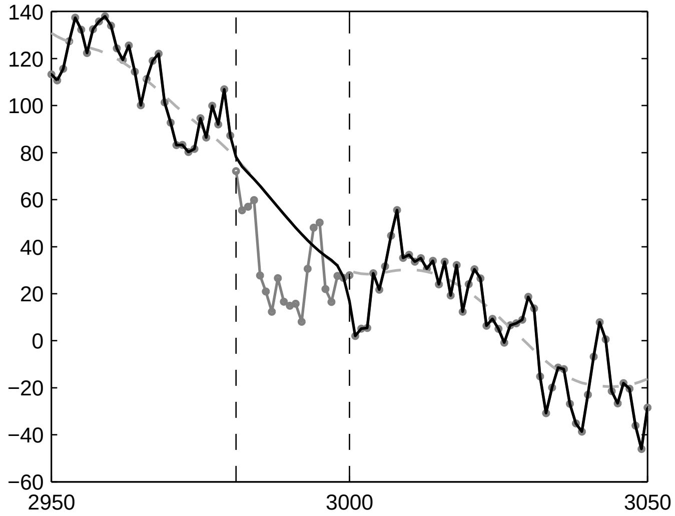

Figure 4.36: The improved prediction over missing data at 2981 – 3000.

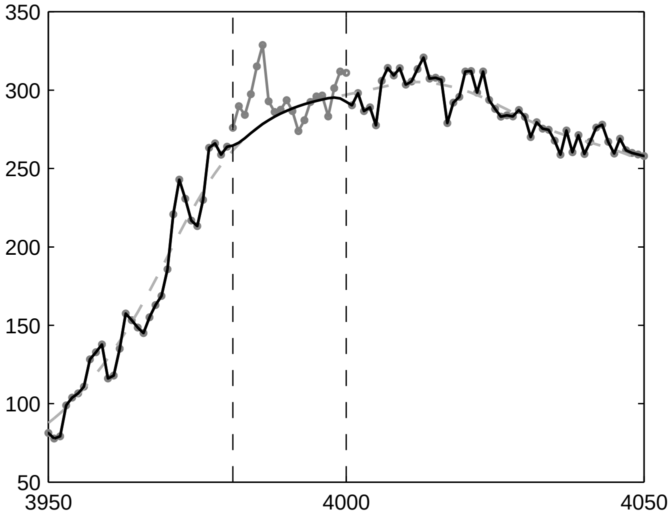

Figure 4.37: The improved prediction over missing data at 3981 – 4000.

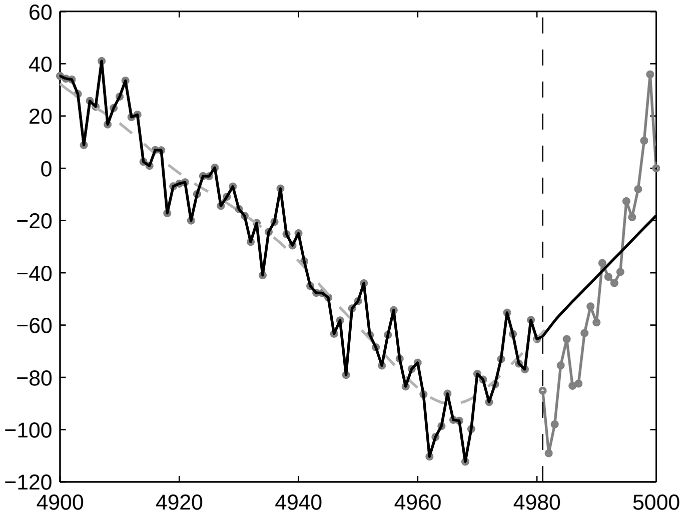

Figure 4.38: The improved prediction over missing data at 4981 – 5000.

Chapter 5

Conclusions and Discussion

5.1 Conclusions

This thesis has presented new algorithms for non-linear and non-Gaussian continuous-discrete (Bayesian) optimal filtering and smoothing, that is, for recursive Bayesian estimation of states of stochastic differential equations, which are observed through discrete-time measurements. The classical theories of discrete-time and continuous-discrete-time optimal filtering have been first presented in Bayesian terms and new algorithms have been developed to overcome limitations of the classical continuous-discrete methods. New algorithms have also been developed for the continuous-time filtering models.

The novel continuous-discrete unscented Kalman filter and smoother are continuous-discrete versions of the unscented Kalman filter and smoother, where the continuous-time prediction is performed using a novel continuous-time unscented transform based approximation method for stochastic differential equations. In this thesis, the equations for the new filter and smoother have been derived and using simulated data, the performance of them has been experimentally compared to other filters and smoothers. Also the equations of the novel continuous-time Kalman-Bucy filter have been derived and experimentally compared to the extended Kalman-Bucy filter.

The continuous-discrete unscented filter and smoother are good alternatives to the extended Kalman filter and smoother in models, where the Jacobian and Hessian matrices of the drift terms are not available. Usually, the estimation performance is quite much the same with both the approaches, but with some models the unscented Kalman filters and smoothers give better results (and in some cases, not). However, in addition to the availability of derivatives, according to the simulation results there does not seem to be clear general guidelines on which should be chosen for a given filtering problem.

This thesis has also presented new class of methods for continuous-discrete particle filtering and smoothing. These methods are based on transformations of probability measures by the Girsanov theorem. The new methods are applicable to a general class of models, in particular, they can be applied to many models with singular dispersion matrices, unlike many previously proposed measure transformation based sampling methods. The new methods have been illustrated in simulated problems, where both the implementation details of the algorithms and the simulation results have been reported. The methods have also been applied to estimation of the spread of an infectious disease based on counts of dead individuals.

The new continuous-discrete unscented Kalman filter as well as the classical continuous-discrete extended Kalman filter can be used for forming importance processes for the new continuous-discrete particle filters. This way the efficiency of the Gaussian approximation based filters can be combined with the accuracy of the particle approximations. Closed form marginalization or Rao-Blackwellization can be applied if the model is conditionally Gaussian or if the model contains unknown static parameters and has a suitable conjugate form. In most cases Rao-Blackwellization leads to significant improvement in efficiency of the particle filtering algorithm.

This thesis has also presented a new particle filtering based algorithm for tracking an unknown number of targets. The algorithm is based on constructing a Rao-Blackwellized particle filter for the probabilistic model constructed for the target states, the data associations and the birth and death processes. Simulations have been used for demonstrating the performance of the algorithm. It seems that particle filtering is well suited to multiple target tracking, because it allows flexible modeling of data associations, appearances and disappearance of the targets. Rao-Blackwellization significantly reduces the dimensionality of the parameter space that needs to be sampled and improves the efficiency of the particle filter. Approximate Rao-Blackwellization with EKF or UKF can be used when the single target tracking sub-problem is solvable with EKF or UKF.

In this thesis, also a new stochastic version of the SIR model for modeling the spread of infections diseases is presented. This model is used as the dynamic model in a Rao-Blackwellized particle filter, which is used for estimating the spread of the disease and its parameters from measured number of deaths or infected individuals. The performance of the algorithm is demonstrated using the classical Bombay plague data. In this case, the particle based methods have the advantage that it is easy to compute predictive quantities such as estimates of the time when the maximum of the epidemic will be reached and estimates of the expected total number of deaths. Due to the recursive nature of the algorithm, the estimates can be computed at each day of the epidemic and the estimates can be updated recursively when new data arrives without need for performing all the previous computations again.

The continuous-discrete Kalman filter and smoother have been applied to the

CATS benchmark time series data with very good results. The original solution, which won the CATS competition that was organized as a special session of the IJCNN 2004 conference and a slightly improved solution have been presented in this thesis. These results show that Kalman filtering based methods, even though they are linear methods, can sometimes be used for modeling non-linear looking time series and the results can be better than the results of far more complicated non-linear methods.

5.2 Practical Considerations

Although in theory particle filters with sufficient number of particles should always be better than Gaussian approximation based methods in non-linear filtering problems, the classical methods should also always be tested. The general guideline is that the extended Kalman filter should always be applied to the filtering problem first. It is almost always possible to approximate and linearize the filtering model such that EKF or a similar algorithm can be applied. If not, some of the model parameters can be assumed to be known and the EKF can be tested with simulated data.

The unscented Kalman filter can also be used in place of the EKF and it has the advantage that it can be implemented without computing the symbolic Jacobian (and Hessian matrices) of the drift and measurement model functions. The UKF solution can also be used for validating (debugging) the EKF implementation, especially if the derivative computations are very complicated and thus error-prone. The approximation of the UKF is closer to the second order EKF than to the first order EKF and for this reason it is a good alternative to EKF in cases, where a second order approximation would be needed but the Hessian matrices are very hard to compute.

When the EKF/UKF solution works, it is a good idea to implement a bootstrap filter based solution. If the EKF/UKF based solution has been implemented to a simplified model, where some of the parameters are assumed to be known, the bootstrap filter can also be implemented to this simplified model first. The bootstrap filter with a sufficient number of particles should give results that are comparable or better than the results of the EKF/UKF. At least the results of the bootstrap should not be significantly worse.

If the model is conditionally Gaussian or if there are unknown static parameters, where Rao-Blackwellization can be applied, this can be done next. Implementing the Rao-Blackwellization to the bootstrap filter should not be a difficult task. It is good idea to Rao-Blackwellize as many parameters as possible because closed form computations are always more efficient than sampling.

When the bootstrap filter works, a better importance process can be constructed with help from the EKF/UKF based solution. At this stage also other possible enhancements to the particle filter can be applied. The results after the enhancements should be the same as with the bootstrap filter, but less particles should be needed for the same accuracy. At this stage it is wise to check that the final particle filter solution is also in practice better than the EKF/UKF solution. Sometimes this is not the case.

5.3 Future Work

As in the discrete-time case, the explicit covariance forms of the continuous-discrete and continuous-time unscented Kalman filters and smoothers are not very stable numerically. The numerical stability could be improved by using square roots of the covariances instead of the plain covariances. This could be achieved by formulating the differential equations of the filters in terms of the sigma-points instead of the means and covariances.

In the filtering models of this thesis the dispersion matrix $\mathbf{L}(t)$ is allowed to depend on time only, and not on the state, that is, the dispersion matrix cannot be of the general form $\mathbf{L}(\mathbf{x}, t)$. It would be possible to generalize the continuous-discrete (and continuous-time) unscented Kalman filter equations to allow the more general dispersion matrices. This would allow modeling of, for example, exponential processes, which are common in financial applications. However, in many cases it is possible to convert a model with state dependent dispersion matrix into a model with state independent dispersion matrix by a suitable change of variables.

In this thesis the importance processes used in continuous-discrete particle filtering examples are very simple and better alternatives definitely exists. In principle, the optimal importance process in the continuous-discrete particle filtering case would have the same law as the smoothing solution. Thus, constructing the importance process based on the smoothing solution instead of linearly interpolated filtering solutions, as in this thesis, could lead to more efficient particle filtering methods. In some cases it could be possible to construct a process, which would have exactly the same law as the optimal importance process.

A weakness in the continuous-discrete particle filtering framework is that the importance process has to be scaled before sampling. In practice, this restricts the possible forms of importance processes to those having the same dispersion matrix as the original process. It could be possible to modify the equations such the scaling of the importance process would not be needed.

The continuous-discrete particle filtering framework could be also extended to cover the case of time dependent dispersion matrix $\mathbf{L}(t)$. This extension should be quite straightforward, because the Girsanov theorem can be applied also in this time-varying case.

The continuous-discrete sequential importance resampling framework could be extended to the case of stochastic differential equations driven by more general martingales, for example, general Lévy processes such as compound Poisson processes. This would allow modeling of sudden changes in signals. This extension could be possible by simply replacing the Brownian motion in the Girsanov theorem by a more general martingale.

It could be possible to generalize the continuous-discrete sequential importance sampling framework presented in this thesis to continuous-time filtering problems. Then the extended Kalman-Bucy filter or the unscented Kalman-Bucy filter could be used for forming the importance process and the actual filtering result would be formed by weighting the importance process samples properly.

The contributions of this thesis are more on the theoretical side than on the practical side of optimal filtering and for this reason many of the experimental data in the thesis are simulated. In the future it would be interesting to see more applications of the algorithms to real world problems. Comparisons to alternative approaches would be also interesting, even with simulated data.

Appendix A

Additional Material

A.1 Properties of Gaussian Distribution

Definition A.1 (Gaussian distribution)**.** *Random variable* $\mathbf{x} \in \mathbb{R}^n$ *has Gaussian distribution with mean* $\mathbf{m} \in \mathbb{R}^n$ *and covariance* $\mathbf{P} \in \mathbb{R}^{n \times n}$ *if it has the probability density of the form*

$$\mathrm{N}(\mathbf{x} \,|\, \mathbf{m}, \mathbf{P}) = \frac{1}{(2\,\pi)^{n/2}\,|\mathbf{P}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{m})^T\,\mathbf{P}^{-1}\,(\mathbf{x}-\mathbf{m})\right), \tag{A.1}$$

where $|\mathbf{P}|$ *is the determinant of matrix* $\mathbf{P}$.

Lemma A.1 (Joint density of Gaussian variables)**.** *If random variables* $\mathbf{x} \in \mathbb{R}^n$ *and* $\mathbf{y} \in \mathbb{R}^m$ *have the Gaussian probability densities*

$$\begin{aligned} \mathbf{x} &\sim \mathrm{N}(\mathbf{x} \,|\, \mathbf{m}, \mathbf{P}) \\ \mathbf{y} \,|\, \mathbf{x} &\sim \mathrm{N}(\mathbf{y} \,|\mathbf{H}\,\mathbf{x} + \mathbf{u}, \mathbf{R}), \end{aligned} \tag{A.2}$$

then the joint density of $\mathbf{x}, \mathbf{y}$ *and the marginal distribution of* $\mathbf{y}$ *are given as*

$$\begin{aligned} \mathbf{x}, \mathbf{y} &\sim \mathrm{N}\left(\begin{bmatrix} \mathbf{m} \\ \mathbf{H}\,\mathbf{m} + \mathbf{u} \end{bmatrix}, \begin{bmatrix} \mathbf{P} & \mathbf{P}\,\mathbf{H}^T \\ \mathbf{H}\,\mathbf{P} & \mathbf{H}\,\mathbf{P}\,\mathbf{H}^T + \mathbf{R} \end{bmatrix}\right) \\ \mathbf{y} &\sim \mathrm{N}(\mathbf{H}\,\mathbf{m} + \mathbf{u}, \mathbf{H}\,\mathbf{P}\,\mathbf{H}^T + \mathbf{R}). \end{aligned} \tag{A.3}$$

Lemma A.2 (Conditional density of Gaussian variables)**.** *If the random variables* $\mathbf{x}$ *and* $\mathbf{y}$ *have the joint Gaussian probability density*

$$\mathbf{x}, \mathbf{y} \sim \mathrm{N}\left(\begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}, \begin{bmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{C}^T & \mathbf{B} \end{bmatrix}\right), \tag{A.4}$$

then the marginal and conditional densities of $\mathbf{x}$ *and* $\mathbf{y}$ *are given as follows:*

$$\begin{aligned} \mathbf{x} &\sim \mathrm{N}(\mathbf{a}, \mathbf{A}) \\ \mathbf{y} &\sim \mathrm{N}(\mathbf{b}, \mathbf{B}) \\ \mathbf{x} \,|\, \mathbf{y} &\sim \mathrm{N}(\mathbf{a} + \mathbf{C}\,\mathbf{B}^{-1}\,(\mathbf{y} - \mathbf{b}), \mathbf{A} - \mathbf{C}\,\mathbf{B}^{-1}\mathbf{C}^T) \\ \mathbf{y} \,|\, \mathbf{x} &\sim \mathrm{N}(\mathbf{b} + \mathbf{C}^T\,\mathbf{A}^{-1}\,(\mathbf{x} - \mathbf{a}), \mathbf{B} - \mathbf{C}^T\,\mathbf{A}^{-1}\,\mathbf{C}). \end{aligned} \tag{A.5}$$

A.2 Cholesky factorization

The Cholesky factor of the symmetric positive definite matrix $\mathbf{P}$ is a lower triangular matrix $\mathbf{A}$ such that

$$\mathbf{P} = \mathbf{A}\,\mathbf{A}^T. \tag{A.6}$$

The matrix $\mathbf{A}$ can be computed by the Cholesky factorization algorithm (see, e.g., Golub and van Loan, 1996) presented below.

Algorithm A.1 (Cholesky factorization)**.** *The Cholesky factor* $\mathbf{A}$ *of matrix* $\mathbf{P}$ *can be computed as follows:*

```
procedure CHOL($\mathbf{P}$)
    for $i \leftarrow 1 \ldots n$ do
        $A_{ii} = \sqrt{P_{ii} - \sum_{k<i} A_{ik}^2}$
        for $j \leftarrow i + 1 \ldots n$ do
            $A_{ji} = \left(P_{ji} - \sum_{k<i} A_{jk}\,A_{ik}\right)/A_{ii}$
        end for
    end for
    return $\mathbf{A}$
end procedure
```

A.3 Derivations of Filters and Smoothers

A.3.1 Derivation of Kalman-Bucy Filter

In this section the equations of the Kalman-Bucy filter are derived. The equations are stochastic differential equations for the mean and covariance of the Gaussian filtering distribution of the linear Gaussian continuous-time filtering model (3.315).

The linear state dynamics

$$\mathrm{d}\mathbf{x} = \mathbf{F}(t)\,\mathbf{x}\,\mathrm{d}t + \mathbf{L}(t)\,\mathrm{d}\boldsymbol{\beta}, \tag{A.7}$$

where $\boldsymbol{\beta}(t)$ is a Brownian motion with diffusion matrix $\mathbf{Q}_c(t)$ can be discretely approximated as

$$\mathbf{x}_k = \mathbf{x}_{k-1} + \mathbf{F}\,\mathbf{x}_{k-1}\,\delta t + \mathbf{q}_{k-1}, \tag{A.8}$$

where $\mathbf{q}_{k-1} \sim \mathrm{N}(\mathbf{0}, \mathbf{L}\,\mathbf{Q}_c\,\mathbf{L}^T\,\delta t)$. The Kalman filter prediction equations for this model are

$$\begin{aligned} \mathbf{m}_k^- &= \mathbf{m}_{k-1} + \mathbf{F}\,\mathbf{m}\,\delta t \\ \mathbf{P}_k^- &= (\mathbf{I} + \mathbf{F}\,\delta t)\,\mathbf{P}_{k-1}\,(\mathbf{I} + \mathbf{F}\,\delta t)^T + \mathbf{L}\,\mathbf{Q}_c\,\mathbf{L}^T\,\delta t \\ &\approx \mathbf{P}_{k-1} + \mathbf{F}\,\mathbf{P}_{k-1}\,\delta t + \mathbf{P}_{k-1}\,\mathbf{F}^T\,\delta t + \mathbf{L}\,\mathbf{Q}_c\,\mathbf{L}^T\,\delta t, \end{aligned} \tag{A.9}$$

where we have retained only the first order terms. The measurement model

$$\mathrm{d}\mathbf{y} = \mathbf{H}(t)\,\mathbf{x}\,\mathrm{d}t + \mathbf{V}(t)\,\mathrm{d}\boldsymbol{\eta}, \tag{A.10}$$

where $\boldsymbol{\eta}(t)$ is a Brownian motion with diffusion matrix $\mathbf{R}_c(t)$ can be similarly approximated as

$$\mathbf{z}\,\delta t = \mathbf{H}\,\mathbf{x}\,\delta t + \mathbf{v}, \tag{A.11}$$

where $\mathbf{z}(t) = \mathrm{d}\mathbf{y}(t)/\mathrm{d}t$ and $\mathbf{v} \sim \mathrm{N}(\mathbf{0}, \mathbf{V}\,\mathbf{R}_c\,\mathbf{V}^T\,\delta t)$. The corresponding Kalman filter update equations are

$$\begin{aligned} \mathbf{S}_k &= \mathbf{H}\,\mathbf{P}_k^-\,\mathbf{H}^T\,\delta t^2 + \mathbf{V}\,\mathbf{R}_c\,\mathbf{V}^T\,\delta t \\ \mathbf{K}_k &= \mathbf{P}_k^-\,\mathbf{H}^T\,\delta t\,\mathbf{S}_k^{-1} \\ \mathbf{m}_k &= \mathbf{m}_k^- + \mathbf{K}_k\,\left[\mathbf{z}\,\delta t - \mathbf{H}\,\delta t\,\mathbf{m}_k^-\right] \\ \mathbf{P}_k &= \mathbf{P}_k^- - \mathbf{K}_k\,\mathbf{S}_k\,\mathbf{K}_k^T. \end{aligned} \tag{A.12}$$

The Kalman gain can be written as

$$\begin{aligned} \mathbf{K}_k &= \mathbf{P}_k^-\,\mathbf{H}^T\,\delta t\,\mathbf{S}_k^{-1} \\ &= \mathbf{P}_k^-\,\mathbf{H}^T\,\delta t\,(\mathbf{H}\,\mathbf{P}_k^-\,\mathbf{H}^T\,\delta t^2 + \mathbf{V}\,\mathbf{R}_c\,\mathbf{V}^T\,\delta t)^{-1} \\ &= \mathbf{P}_k^-\,\mathbf{H}^T\,(\mathbf{H}\,\mathbf{P}_k^-\,\mathbf{H}^T\,\delta t + \mathbf{V}\,\mathbf{R}_c\,\mathbf{V}^T)^{-1}, \end{aligned} \tag{A.13}$$

and the second term in covariance update can be written as

$$\begin{aligned} \mathbf{K}_k\,\mathbf{S}_k\,\mathbf{K}_k^T &= \mathbf{P}_k^-\,\mathbf{H}^T\,\delta t\,(\mathbf{H}\,\mathbf{P}_k^-\,\mathbf{H}^T\,\delta t^2 + \mathbf{V}\,\mathbf{R}_c\,\mathbf{V}^T\,\delta t)^{-1}\,\delta t\,\mathbf{P}_k^-\,\mathbf{H} \\ &= \mathbf{P}_k^-\,\mathbf{H}^T\,(\mathbf{H}\,\mathbf{P}_k^-\,\mathbf{H}^T\,\delta t + \mathbf{V}\,\mathbf{R}_c\,\mathbf{V}^T)^{-1}\,\mathbf{P}_k^-\,\mathbf{H}\,\delta t \\ &= \mathbf{K}_k\,(\mathbf{H}\,\mathbf{P}_k^-\,\mathbf{H}^T\,\delta t + \mathbf{V}\,\mathbf{R}_c\,\mathbf{V}^T)\mathbf{K}_k^T\,\delta t. \end{aligned} \tag{A.14}$$

If we now substitute the dynamic model we get

$$\begin{aligned} \mathbf{m}_k &= \mathbf{m}_{k-1} + \mathbf{F}\,\mathbf{m}\,\delta t + \mathbf{K}_k\,\left[\mathbf{z} - \mathbf{H}\,\mathbf{m}_k^-\right]\,\delta t \\ \mathbf{P}_k &= \mathbf{P}_{k-1} + \mathbf{F}\,\mathbf{P}_{k-1}\,\delta t + \mathbf{P}_{k-1}\,\mathbf{F}^T\,\delta t + \mathbf{L}\,\mathbf{Q}_c\,\mathbf{L}^T\,\delta t \\ &\quad - \mathbf{K}_k\,(\mathbf{H}\,\mathbf{P}_k^-\,\mathbf{H}^T\,\delta t + \mathbf{V}\,\mathbf{R}_c\,\mathbf{V}^T)\mathbf{K}_k^T\,\delta t. \end{aligned} \tag{A.15}$$

By rearranging

$$\begin{aligned}\frac{\mathbf{m}_k - \mathbf{m}_{k-1}}{\delta t} &= \mathbf{F}\,\mathbf{m} + \mathbf{K}_k\,\left[\mathbf{z} - \mathbf{H}\,\mathbf{m}_k^-\right] \\ \frac{\mathbf{P}_k - \mathbf{P}_{k-1}}{\delta t} &= \mathbf{F}\,\mathbf{P}_{k-1} + \mathbf{P}_{k-1}\,\mathbf{F}^T + \mathbf{L}\,\mathbf{Q}_c\,\mathbf{L}^T \\ &- \mathbf{K}_k\,(\mathbf{H}\,\mathbf{P}_k^-\,\mathbf{H}^T\,\delta t + \mathbf{V}\,\mathbf{R}_c\,\mathbf{V}^T)\mathbf{K}_k^T.\end{aligned} \tag{A.16}$$

In the limit $\delta t \to 0$ we get the result in the Theorem 3.8.

Another way of deriving the Kalman-Bucy filter equations is to convert the Wiener-Hopf equation (Wiener, 1950) into differential equations for the sufficient statistics (see, e.g., Van Trees, 1968). That is the derivation that was presented in the original article of Kalman and Bucy (1961).

A.3.2 Derivation of Kalman-Bucy Smoother

In this section we derive the optimal smoother to the filtering problems where the state dynamics are linear Gaussian and the filtering distributions are Gaussian. This kind of filtering problems are the continuous-discrete Kalman filtering problem and the continuous-time Kalman-Bucy filtering problem. Thus, the optimal smoother is the same for both of these filtering problems.

For the discretized dynamic model

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{F}(t_k)\,\mathbf{x}_k\,\delta t + \mathbf{q}_k, \tag{A.17}$$

where $\mathbf{q}_k \sim \mathrm{N}(\mathbf{0}, \mathbf{L}(t_k)\,\mathbf{Q}(t_k)\,\mathbf{L}^T(t_k)\,\delta t)$, the first two discrete-time smoothing equations can be written up to first order in δt as

$$\begin{aligned}\mathbf{m}_{k+1}^- &= \mathbf{m}_k + \mathbf{F}(t_k)\,\mathbf{m}_k\,\delta t \\ \mathbf{P}_{k+1}^- &= (\mathbf{I} + \mathbf{F}(t_k)\,\delta t)\,\mathbf{P}_k\,(\mathbf{I} + \mathbf{F}(t_k)\,\delta t)^T + \mathbf{L}(t_k)\,\mathbf{Q}(t_k)\,\mathbf{L}^T(t_k)\,\delta t \\ &\approx \mathbf{P}_k + \mathbf{F}(t_k)\,\mathbf{P}_k\,\delta t + \mathbf{P}_k\,\mathbf{F}^T(t_k)\,\delta t + \mathbf{L}(t_k)\,\mathbf{Q}(t_k)\,\mathbf{L}^T(t_k)\,\delta t.\end{aligned} \tag{A.18}$$

By the normal differentiation rules, the differential of the gain (when P is fixed) can be written as

$$\partial[\mathbf{C}_k] = \mathbf{P}_k\,\partial\mathbf{A}_k^T\,\mathbf{P}_k^{-1} - \mathbf{P}_k\,\mathbf{P}_k^{-1}\,\partial\mathbf{P}_{k+1}^-\,\mathbf{P}_k^{-1}, \tag{A.19}$$

and thus up to first order in δt the gain is

$$\begin{aligned}\mathbf{C}_k &\approx \mathbf{I} + \mathbf{P}_k\,\mathbf{F}(t_k)\,\mathbf{P}_k^{-1}\,\delta t - \mathbf{P}_k\,\mathbf{P}_k^{-1}\,[\mathbf{F}(t_k)\,\mathbf{P}_k\,\delta t \\ &\quad + \mathbf{P}_k\,\mathbf{F}^T(t_k)\,\delta t + \mathbf{L}(t_k)\,\mathbf{Q}(t_k)\,\mathbf{L}^T(t_k)\,\delta t]\,\mathbf{P}_k^{-1} \\ &= \mathbf{I} + \mathbf{P}_k\,\mathbf{F}(t_k)\,\mathbf{P}_k^{-1}\,\delta t \\ &\quad - [\mathbf{F}(t_k)\,\mathbf{P}_k\,\delta t + \mathbf{P}_k\,\mathbf{F}^T(t_k)\,\delta t + \mathbf{L}(t_k)\,\mathbf{Q}(t_k)\,\mathbf{L}^T(t_k)\,\delta t]\mathbf{P}_k^{-1} \\ &= \mathbf{I} - \mathbf{F}(t_k)\,\delta t - \mathbf{L}(t_k)\,\mathbf{Q}(t_k)\,\mathbf{L}^T(t_k)\mathbf{P}_k^{-1}\,\delta t.\end{aligned} \tag{A.20}$$

The smoothed mean is given as

$$\begin{aligned}
\mathbf{m}_k^s &= \mathbf{m}_k + \mathbf{C}_k \, [\mathbf{m}_{k+1}^s - \mathbf{m}_{k+1}^-] \\
&\approx \mathbf{m}_k + (\mathbf{I} - \mathbf{F}(t_k) \, \delta t - \mathbf{L}(t_k) \, \mathbf{Q}(t_k) \, \mathbf{L}^T(t_k) \, \mathbf{P}_k^{-1} \, Dt) \\
&\quad \times [\mathbf{m}_{k+1}^s - \mathbf{m}_k - \mathbf{F}(t_k) \, \mathbf{m}_k \, \delta t] \\
&\approx \mathbf{m}_{k+1}^s - \mathbf{F}(t_k) \, \mathbf{m}_{k+1}^s \, \delta t - \mathbf{L}(t_k) \, \mathbf{Q}(t_k) \, \mathbf{L}^T(t_k) \, \mathbf{P}_k^{-1} \, \mathbf{m}_{k+1}^s \, \delta t \\
&\quad + \mathbf{L}(t_k) \, \mathbf{Q}(t_k) \, \mathbf{L}^T(t_k) \, \mathbf{P}_k^{-1} \, \mathbf{m}_k \, \delta t,
\end{aligned} \tag{A.21}$$

where we have again retained only the first order terms. The smoothed covariance is given up to first order as

$$\begin{aligned}
\mathbf{P}_k^s &= \mathbf{P}_k + \mathbf{C}_k \, [\mathbf{P}_{k+1}^s - \mathbf{P}_{k+1}^-] \, \mathbf{C}_k^T \\
&\approx \mathbf{P}_k + (\mathbf{I} - \mathbf{F}(t_k) \, \delta t - \mathbf{L}(t_k) \, \mathbf{Q}(t_k) \, \mathbf{L}^T(t_k) \, \mathbf{P}_k^{-1} \, \delta t) \\
&\quad \times [\mathbf{P}_{k+1}^s - \mathbf{P}_k - \mathbf{F}(t_k) \, \mathbf{P}_k \, \delta t - \mathbf{P}_k \, \mathbf{F}^T(t_k) \, \delta t - \mathbf{L}(t_k) \, \mathbf{Q}(t_k) \, \mathbf{L}^T(t_k) \, \delta t] \\
&\quad \times (\mathbf{I} - \mathbf{F}(t_k) \, \delta t - \mathbf{L}(t_k) \, \mathbf{Q}(t_k) \, \mathbf{L}^T(t_k) \, \mathbf{P}_k^{-1} \, \delta t)^T \\
&\approx \mathbf{P}_{k+1}^s - \mathbf{F}(t_k) \, \mathbf{P}_{k+1}^s \, \delta t - \mathbf{L}(t_k) \, \mathbf{Q}(t_k) \, \mathbf{L}^T(t_k) \, \mathbf{P}_k^{-1} \, \mathbf{P}_{k+1}^s \, \delta t \\
&\quad - \mathbf{P}_{k+1}^s \, \mathbf{F}^T(t_k) \, \delta t - \mathbf{P}_{k+1}^s \, \mathbf{P}_k^{-1} \, \mathbf{L}(t_k) \, \mathbf{Q}(t_k) \, \mathbf{L}^T(t_k) \, \delta t \\
&\quad + \mathbf{L}(t_k) \, \mathbf{Q}(t_k) \, \mathbf{L}^T(t_k) \, \delta t.
\end{aligned} \tag{A.22}$$

Rearranging and dividing by δt gives

$$\begin{aligned}
\frac{\mathbf{m}_{k+1}^s - \mathbf{m}_k^s}{\delta t} &= \mathbf{F}(t_k) \, \mathbf{m}_{k+1}^s + \mathbf{L}(t_k) \, \mathbf{Q}(t_k) \, \mathbf{L}^T(t_k) \, \mathbf{P}_k^{-1} \, \mathbf{m}_{k+1}^s \\
&\qquad - \mathbf{L}(t_k) \, \mathbf{Q}(t_k) \, \mathbf{L}^T(t_k) \, \mathbf{P}_k^{-1} \, m_k \\
\frac{\mathbf{P}_{k+1}^s - \mathbf{P}_k^s}{\delta t} &= \mathbf{F}(t_k) \, \mathbf{P}_{k+1}^s + \mathbf{L}(t_k) \, \mathbf{Q}(t_k) \, \mathbf{L}^T(t_k) \, \mathbf{P}_k^{-1} \, \mathbf{P}_{k+1}^s + \mathbf{P}_{k+1}^s \, \mathbf{F}^T(t_k) \\
&\qquad + \mathbf{P}_{k+1}^s \, \mathbf{L}(t_k) \, \mathbf{Q}(t_k) \, \mathbf{L}^T(t_k) \, \mathbf{P}_k^{-1} - \mathbf{L}(t_k) \, \mathbf{Q}(t_k) \, \mathbf{L}^T(t_k),
\end{aligned} \tag{A.23}$$

and in the limit $\delta t \to 0$ we get the result in the Algorithm 3.18.

A.3.3 Derivation of Unscented Kalman-Bucy Filter

In this section we give the details of derivation of the unscented Kalman-Bucy filter equations in Theorem 3.9, which are the differential equations for the unscented transform based approximation to the mean and covariance of the filtering distribution of the non-linear continuous-time filtering model. As opposed to the derivation of Kalman-Bucy filter in Section A.3.1 this derivation explicitly uses the integral equation interpretation of the Itô stochastic differential equations in the continuous-time filtering model.

The continuous-time filtering model

$$\begin{aligned} \mathrm{d}\mathbf{x}(t) &= \mathbf{f}(\mathbf{x}(t), t)\,\mathrm{d}t + \mathbf{L}(t)\,\mathrm{d}\boldsymbol{\beta}(t) \\ \mathrm{d}\mathbf{y}(t) &= \mathbf{h}(\mathbf{x}(t), t)\,\mathrm{d}t + \mathbf{V}(t)\,\mathrm{d}\boldsymbol{\eta}(t), \end{aligned} \tag{A.24}$$

where $\boldsymbol{\beta}(t)$ and $\boldsymbol{\eta}(t)$ are independent Brownian motions with diagonal diffusion matrices $\mathbf{Q}_c(t)$ and $\mathbf{R}_c(t)$ can be interpreted as a pair of stochastic integral equations such that for all $\delta t \geq 0$ we have

$$\mathbf{x}(t + \delta t) - \mathbf{x}(t) = \int_t^{t+\delta t} \mathbf{f}(\mathbf{x}(t), t)\,\mathrm{d}t + \int_t^{t+\delta t} \mathbf{L}(t)\,\mathrm{d}\boldsymbol{\beta}(t) \tag{A.25}$$

$$\mathbf{y}(t + \delta t) - \mathbf{y}(t) = \int_t^{t+\delta t} \mathbf{h}(\mathbf{x}(t), t)\,\mathrm{d}t + \int_t^{t+\delta t} \mathbf{V}(t)\,\mathrm{d}\boldsymbol{\eta}(t). \tag{A.26}$$

For given δt the dynamic model (A.25) is discrete and the UKF prediction step can be written as[1]

$$\mathbf{A}(t) = \mathrm{chol}(\mathbf{P}(t)) \tag{A.27}$$

$$\begin{aligned} \mathbf{X}(t) &= \begin{bmatrix} \mathbf{m}(t) & \cdots & \mathbf{m}(t) \end{bmatrix} \\ &\quad + \sqrt{c} \begin{bmatrix} \mathbf{0} & \mathbf{A}(t) & -\mathbf{A}(t) \end{bmatrix} \end{aligned} \tag{A.28}$$

$$\tilde{\mathbf{X}}(t + \delta t) = \mathbf{X}(t) + \int_t^{t+\delta t} \mathbf{f}(\mathbf{X}(t), t)\,\mathrm{d}t \tag{A.29}$$

$$\mathbf{m}^-(t + \delta t) = \tilde{\mathbf{X}}(t + \delta t)\,\mathbf{w}_m \tag{A.30}$$

$$\mathbf{P}^-(t + \delta t) = \tilde{\mathbf{X}}(t + \delta t)\,\mathbf{W}\,\tilde{\mathbf{X}}^T(t + \delta t) + \int_t^{t+\delta t} \mathbf{L}(t)\,\mathbf{Q}_c(t)\,\mathbf{L}^T(t)\,\mathrm{d}t. \tag{A.31}$$

Eliminating the sigma points $\tilde{\mathbf{X}}(t + \delta t)$, from equations (A.27) – (A.31) gives

$$\begin{aligned} \mathbf{m}^-(t + \delta t) &= \mathbf{X}(t)\,\mathbf{w}_m + \left[\int_t^{t+\delta t} \mathbf{f}(\mathbf{X}(t), t)\,\mathrm{d}t \right] \mathbf{w}_m \\ &= \mathbf{m}(t) + \left[\int_t^{t+\delta t} \mathbf{f}(\mathbf{X}(t), t)\,\mathrm{d}t \right] \mathbf{w}_m, \end{aligned} \tag{A.32}$$

[1] Note that the equations should be interpreted as implicit integral equations for the solution, because the integrals of $\mathbf{X}(t)$ cannot be evaluated before knowing $\mathbf{X}(t)$.

and

$$\begin{aligned}\mathbf{P}^-(t+\delta t) &= \left[\mathbf{X}(t) + \int_t^{t+\delta t} \mathbf{f}(\mathbf{X}(t),t)\,\mathrm{d}t\right] \mathbf{W} \left[\mathbf{X}(t) + \int_t^{t+\delta t} \mathbf{f}(\mathbf{X}(t),t)\,\mathrm{d}t\right]^T \\ &\quad + \int_t^{t+\delta t} \mathbf{L}(t)\,\mathbf{Q}_c(t)\,\mathbf{L}^T(t)\,\mathrm{d}t \\ &= \mathbf{P}(t) + \mathbf{X}(t)\,\mathbf{W}\,\left[\int_t^{t+\delta t} \mathbf{f}(\mathbf{X}(t),t)\,\mathrm{d}t\right]^T \\ &\quad + \left[\int_t^{t+\delta t} \mathbf{f}(\mathbf{X}(t),t)\,\mathrm{d}t\right]\,\mathbf{W}\,\mathbf{X}^T(t) \\ &\quad + \left[\int_t^{t+\delta t} \mathbf{f}(\mathbf{X}(t),t)\,\mathrm{d}t\right]\,\mathbf{W}\,\left[\int_t^{t+\delta t} \mathbf{f}(\mathbf{X}(t),t)\,\mathrm{d}t\right]^T \\ &\quad + \int_t^{t+\delta t} \mathbf{L}(t)\,\mathbf{Q}_c(t)\,\mathbf{L}^T(t)\,\mathrm{d}t. \end{aligned} \tag{A.33}$$

If we assume that we actually measure the difference

$$\Delta\mathbf{y}(t+\delta t) = \mathbf{y}(t+\delta t) - \mathbf{y}(t), \tag{A.34}$$

the UKF update step can be written as

$$\mathbf{B}(t+\delta t) = \mathrm{chol}(\mathbf{P}^-(t+\delta t)) \tag{A.35}$$

$$\begin{aligned}\mathbf{X}^-(t+\delta t) &= \begin{bmatrix}\mathbf{m}^-(t) & \cdots & \mathbf{m}^-(t)\end{bmatrix} \\ &\quad + \sqrt{c}\,\begin{bmatrix}\mathbf{0} & \mathbf{B}(t+\delta t) & -\mathbf{B}(t+\delta t)\end{bmatrix}\end{aligned} \tag{A.36}$$

$$\Delta\mathbf{Y}(t+\delta t) = \int_t^{t+\delta t} \mathbf{h}(\mathbf{X}^-(t),t)\,\mathrm{d}t \tag{A.37}$$

$$\boldsymbol{\mu}(t+\delta t) = \Delta\mathbf{Y}(t+\delta t)\,\mathbf{w}_m \tag{A.38}$$

$$\begin{aligned}\mathbf{S}(t+\delta t) &= \Delta\mathbf{Y}(t+\delta t)\,\mathbf{W}\,\Delta\mathbf{Y}^T(t+\delta t) \\ &\quad + \int_t^{t+\delta t} \mathbf{V}(t)\,\mathbf{R}_c(t)\,\mathbf{V}^T(t)\,\mathrm{d}t\end{aligned} \tag{A.39}$$

$$\mathbf{C}(t+\delta t) = \mathbf{X}^-(t+\delta t)\,\mathbf{W}\,\Delta\mathbf{Y}^T(t+\delta t) \tag{A.40}$$

$$\mathbf{K}(t+\delta t) = \mathbf{C}(t+\delta t)\,\mathbf{S}^{-1}(t+\delta t) \tag{A.41}$$

$$\begin{aligned}\mathbf{m}(t+\delta t) &= \mathbf{m}^-(t+\delta t) + \mathbf{K}(t+\delta t) \\ &\quad \times [\Delta\mathbf{y}(t+\delta t) - \boldsymbol{\mu}(t+\delta t)]\end{aligned} \tag{A.42}$$

$$\begin{aligned}\mathbf{P}(t+\delta t) &= \mathbf{P}^-(t+\delta t) \\ &\quad - \mathbf{K}(t+\delta t)\,\mathbf{S}(t+\delta t)\,\mathbf{K}^T(t+\delta t).\end{aligned} \tag{A.43}$$

If we eliminate the predicted mean and covariance we get

$$\begin{aligned}\mathbf{m}(t+\delta t) &= \mathbf{m}(t) + \left[\int_t^{t+\delta t} \mathbf{f}(\mathbf{X}(t),t)\,\mathrm{d}t\right]\mathbf{w}_m \\ &\quad + \mathbf{K}(t+\delta t)\left[\Delta\mathbf{y}(t+\delta t) - \Big(\int_t^{t+\delta t} \mathbf{h}(\mathbf{X}^-(t),t)\,\mathrm{d}t\Big)\mathbf{w}_m\right]\end{aligned} \tag{A.44}$$

$$\begin{aligned}\mathbf{P}(t+\delta t) &= \mathbf{P}(t) + \mathbf{X}(t)\,\mathbf{W}\,\left[\int_t^{t+\delta t} \mathbf{f}(\mathbf{X}(t),t)\mathrm{d}t\right]^T \\ &\quad + \left[\int_t^{t+\delta t} \mathbf{f}(\mathbf{X}(t),t)\mathrm{d}t\right]\,\mathbf{W}\,\mathbf{X}^T(t) \\ &\quad + \left[\int_t^{t+\delta t} \mathbf{f}(\mathbf{X}(t),t)\mathrm{d}t\right]\,\mathbf{W}\,\left[\int_t^{t+\delta t} \mathbf{f}(\mathbf{X}(t),t)\mathrm{d}t\right]^T \\ &\quad + \int_t^{t+\delta t} \mathbf{L}(t)\,\mathbf{Q}_c(t)\,\mathbf{L}^T(t)\,\mathrm{d}t \\ &\quad - \mathbf{K}(t+\delta t)\Bigg\{\Big[\int_t^{t+\delta t} \mathbf{h}(\mathbf{X}^-(t),t)\,\mathrm{d}t\Big] \\ &\qquad \times \mathbf{W}\Big[\int_t^{t+\delta t} \mathbf{h}(\mathbf{X}^-(t),t)\,\mathrm{d}t\Big]^T \\ &\quad + \int_t^{t+\delta t} \mathbf{V}(t)\,\mathbf{R}_c(t)\,\mathbf{V}^T(t)\,\mathrm{d}t\Bigg\}\,\mathbf{K}^T(t+\delta t).\end{aligned} \tag{A.45}$$

When δt is small, the drift terms become

$$\int_t^{t+\delta t} \mathbf{f}(\mathbf{X}(t),t)\mathrm{d}t = \mathbf{f}(\mathbf{X}(t),t)\,\delta t + o(\delta t) \tag{A.46}$$

$$\int_t^{t+\delta t} \mathbf{h}(\mathbf{X}^-(t),t)\,\mathrm{d}t = \mathbf{h}(\mathbf{X}^-(t),t)\,\delta t + o(\delta t). \tag{A.47}$$

The following terms vanish completely:

$$\left[\int_t^{t+\delta t} \mathbf{f}(\mathbf{X}(t),t)\mathrm{d}t\right]\,\mathbf{W}\,\left[\int_t^{t+\delta t} \mathbf{f}(\mathbf{X}(t),t)\mathrm{d}t\right]^T = o(\delta t) \tag{A.48}$$

$$\Big[\int_t^{t+\delta t} \mathbf{h}(\mathbf{X}^-(t),t)\,\mathrm{d}t\Big]\,\mathbf{W}\,\Big[\int_t^{t+\delta t} \mathbf{h}(\mathbf{X}^-(t),t)\,\mathrm{d}t\Big]^T = o(\delta t), \tag{A.49}$$

and the diffusion and gain terms become

$$\begin{aligned}
\int_t^{t+\delta t} \mathbf{V}(t)\,\mathbf{R}_c(t)\,\mathbf{V}^T(t)\,\mathrm{d}t &= \mathbf{V}(t)\,\mathbf{R}_c(t)\,\mathbf{V}^T(t)\,\delta t + o(\delta t) && \text{(A.50)} \\
\int_t^{t+\delta t} \mathbf{L}(t)\,\mathbf{Q}_c(t)\,\mathbf{L}^T(t)\,\mathrm{d}t &= \mathbf{L}(t)\,\mathbf{Q}_c(t)\,\mathbf{L}^T(t)\,\delta t + o(\delta t) && \text{(A.51)} \\
\mathbf{K}(t+\delta t) &= \mathbf{X}^-(t+\delta t)\,\mathbf{W}\,\mathbf{h}^T(\mathbf{X}^-(t),t) \\
&\quad \times \left[\mathbf{V}(t)\,\mathbf{R}_c(t)\,\mathbf{V}^T(t)\right]^{-1} + o(\delta t)/\delta t. && \text{(A.52)}
\end{aligned}$$

Substituting these into the mean and covariance equations gives

$$\begin{aligned}
&\mathbf{m}(t+\delta t) \\
&\quad = \mathbf{m}(t) + \mathbf{f}(\mathbf{X}(t),t)\,\mathbf{w}_m\,\delta t \\
&\qquad + \mathbf{K}(t+\delta t)\left[\Delta\mathbf{y}(t+\delta t) - \mathbf{h}(\mathbf{X}^-(t),t)\,\mathbf{w}_m\,\delta t\right] + o(\delta t) && \text{(A.53)} \\
&\mathbf{P}(t+\delta t) \\
&\quad = \mathbf{P}(t) + \mathbf{X}(t)\,\mathbf{W}\,\mathbf{f}^T(\mathbf{X}(t),t)\,\delta t \\
&\qquad + \mathbf{f}(\mathbf{X}(t),t)\,\mathbf{W}\,\mathbf{X}^T(t)\,\delta t + \mathbf{L}(t)\,\mathbf{Q}_c(t)\,\mathbf{L}^T(t)\,\delta t \\
&\qquad - \mathbf{K}(t+\delta t)\,\mathbf{V}(t)\,\mathbf{R}_c(t)\,\mathbf{V}^T(t)\,\mathbf{K}^T(t+\delta t)\,\delta t + o(\delta t). && \text{(A.54)}
\end{aligned}$$

where

$$\Delta\mathbf{y}(t+\delta t) = \mathbf{y}(t+\delta t) - \mathbf{y}(t). \tag{A.55}$$

Rearranging, dividing by δt and taking the limit $\delta t \to 0$ gives the equations (3.321) and (3.320) in the Theorem 3.9. If the stochastic process $\mathbf{y}(t)$ is not differentiable, the limit has to be taken without dividing with δt, and the result is the integral form stochastic differential equation (3.322) for the mean. Note that the equation (3.322) has been rearranged such that the time differential and stochastic differential parts of the equation have been collected together.

A.3.4 Derivation of Unscented Kalman-Bucy Smoother

The continuous-time stochastic differential equation

$$\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x},t)\,\mathrm{d}t + \mathbf{L}(t)\,\mathrm{d}\boldsymbol{\beta}, \tag{A.56}$$

can be approximated up to first order in δt by a discretized dynamic model

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{f}(\mathbf{x}_k,t_k)\,\delta t + \mathbf{q}_k, \tag{A.57}$$

where $\mathbf{q}_k \sim \mathrm{N}(\mathbf{0}, \mathbf{L}(t_k)\,\mathbf{Q}(t_k)\,\mathbf{L}^T(t_k)\,\delta t)$. UKF prediction equations for this model can be written as

$$\begin{aligned}
\mathbf{m}_{k+1}^- &= \mathbf{m}_k + \mathbf{f}(\mathbf{X}_k, t_k)\,\mathbf{w}_m\,\delta t \\
\mathbf{P}_{k+1}^- &= \mathbf{P}_k + \mathbf{f}(\mathbf{X}_k, t_k)\,\mathbf{W}\,\mathbf{X}_k^T\,\delta t + \mathbf{X}_k\,\mathbf{W}\,\mathbf{f}^T(\mathbf{X}_k, t_k)\,\delta t \\
&\qquad + \mathbf{L}(t_k)\,\mathbf{Q}(t_k)\,\mathbf{L}^T(t_k)\,\delta t \\
&= \mathbf{P}_k + \partial\mathbf{P}_k\,\delta t,
\end{aligned}$$

where we have introduced the new variable

$$\partial\mathbf{P}_k = \mathbf{f}(\mathbf{X}_k, t_k)\,\mathbf{W}\,\mathbf{X}_k^T + \mathbf{X}_k\,\mathbf{W}\,\mathbf{f}^T(\mathbf{X}_k, t_k) + \mathbf{L}(t_k)\,\mathbf{Q}(t_k)\,\mathbf{L}^T(t_k), \tag{A.58}$$

and the cross-covariance of $\mathbf{x}_k$ and $\mathbf{x}_{k+1}$ is

$$\begin{aligned}
\mathbf{B}_k^- &= \mathbf{X}_k\,\mathbf{W}\,(\mathbf{X}_k + \mathbf{f}^T(\mathbf{X}_k, t_k)\,\delta t) \\
&= \mathbf{X}_k\,\mathbf{W}\,\mathbf{X}_k + \mathbf{X}_k\,\mathbf{W}\,\mathbf{f}^T(X_k, t_k)\,\delta t \\
&= \mathbf{P}_k + \mathbf{X}_k\,\mathbf{W}\,\mathbf{f}^T(\mathbf{X}_k, t_k)\delta t.
\end{aligned} \tag{A.59}$$

As in the continuous-discrete Kalman smoother case, we may use the differential formula in Equation (A.19), and up to first order in δt the gain is

$$\begin{aligned}
\mathbf{C}_k &\approx \mathbf{P}_k\,\mathbf{P}_k^{-1} + \mathbf{X}_k\,\mathbf{W}\,\mathbf{f}^T(\mathbf{X}_k, t_k)\,\mathbf{P}_{\mathbf{k}}^{-1}\,\delta t - \mathbf{P}_k\,\mathbf{P}_k^{-1}\,[\mathbf{f}(\mathbf{X}_k, t_k)\,\mathbf{W}\,\mathbf{X}_k^T \\
&\qquad + \mathbf{X}_k\,\mathbf{W}\,\mathbf{f}^T(\mathbf{X}_k, t_k) + \mathbf{L}(t_k)\,\mathbf{Q}(t_k)\,\mathbf{L}^T(t_k)]\,\mathbf{P}_k^{-1}\,\delta t \\
&\approx \mathbf{I} + \mathbf{X}_k\,\mathbf{W}\,\mathbf{f}^T(\mathbf{X}_k, t_k)\,\mathbf{P}_k^{-1}\,\delta t \\
&\qquad - \mathbf{f}(\mathbf{X}_k, t_k)\,\mathbf{W}\,\mathbf{X}_k^T\,\mathbf{P}_k^{-1}\,\delta t \\
&\qquad - \mathbf{X}_k\,\mathbf{W}\,\mathbf{f}^T(\mathbf{X}_k, t_k)\,\mathbf{P}_k^{-1}\,\delta t - \mathbf{L}(t_k)\,\mathbf{Q}(t_k)\,\mathbf{L}^T(t_k)]\,\mathbf{P}_k^{-1}\,\delta t \\
&\approx \mathbf{I} - \mathbf{f}(\mathbf{X}_k, t_k)\,\mathbf{W}\,\mathbf{X}_k^T\,\mathbf{P}_k^{-1}\,\delta t - \mathbf{L}(t_k)\,\mathbf{Q}(t_k)\,\mathbf{L}^T(t_k)\,\mathbf{P}_k^{-1}\,\delta t \\
&= \mathbf{I} + \partial\mathbf{C}_k\,\delta t,
\end{aligned} \tag{A.60}$$

where

$$\partial\mathbf{C}_k = -\mathbf{f}(\mathbf{X}_k, t_k)\,\mathbf{W}\,\mathbf{X}_k^T\,\mathbf{P}_k^{-1} - \mathbf{L}(t_k)\,\mathbf{Q}(t_k)\,\mathbf{L}^T(t_k)\,\mathbf{P}_k^{-1}. \tag{A.61}$$

The smoothed mean is given as (up to first order)

$$\begin{aligned}
\mathbf{m}_k^s &= \mathbf{m}_k + \mathbf{C}_k\,[\mathbf{m}_{k+1}^s - \mathbf{m}_{k+1}^-] \\
&= \mathbf{m}_k + \mathbf{C}_k\,[\mathbf{m}_{k+1}^s - \mathbf{m}_k - \mathbf{f}(\mathbf{X}_k, t_k)\,\mathbf{w}_m\,\delta t] \\
&\approx \mathbf{m}_k + (\mathbf{I} + \partial\mathbf{C}_k\,\delta t)[\mathbf{m}_{k+1}^s - \mathbf{m}_k - \mathbf{f}(\mathbf{X}_k, t_k)\,\mathbf{w}_m\,\delta t] \\
&\approx \mathbf{m}_{k+1}^s - \mathbf{f}(\mathbf{X}_k, t_k)\,\mathbf{w}_m\,\delta t + \partial\mathbf{C}_k\,\mathbf{m}_{k+1}^s\,\delta t - \partial\mathbf{C}_k\,\mathbf{m}_k\,\delta t \\
&= \mathbf{m}_{k+1}^s - \mathbf{f}(\mathbf{X}_k, t_k)\,\mathbf{w}_m\,\delta t \\
&\qquad - \mathbf{f}(\mathbf{X}_k, t_k)\,\mathbf{W}\,\mathbf{X}_k^T\,\mathbf{P}_k^{-1}\,\mathbf{m}_{k+1}^s\,\delta t - \mathbf{L}(t_k)\,\mathbf{Q}(t_k)\,\mathbf{L}^T(t_k)\,\mathbf{P}_k^{-1}\,\mathbf{m}_{k+1}^s\,\delta t \\
&\qquad + \mathbf{f}(\mathbf{X}_k, t_k)\,\mathbf{W}\,\mathbf{X}_k^T\,\mathbf{P}_k^{-1}\,\mathbf{m}_k\,\delta t + \mathbf{L}(t_k)\,\mathbf{Q}(t_k)\,\mathbf{L}^T(t_k)\,\mathbf{P}_k^{-1}\,\mathbf{m}_k\,\delta t.
\end{aligned} \tag{A.62}$$

The covariance can be written as (up to first order)

$$
\begin{aligned}
\mathbf{P}_k^s &= \mathbf{P}_k + \mathbf{C}_k \, (\mathbf{P}_{k+1}^s - \mathbf{P}_{k+1}^-) \, \mathbf{C}_k^T \\
&= \mathbf{P}_k + (\mathbf{I} + \partial \mathbf{C}_k \, \delta t) \, (\mathbf{P}_{k+1}^s - \mathbf{P}_k - \partial \mathbf{P}_k \, \delta t) \, (\mathbf{I} + \partial \mathbf{C}_k \, \delta t) \\
&\approx \mathbf{P}_{k+1}^s - \partial \mathbf{P}_k \, \delta t + \partial \mathbf{C}_k \, \mathbf{P}_{k+1}^s \, \delta t - \partial \mathbf{C}_k \, \mathbf{P}_k \, \delta t \\
&\quad + \mathbf{P}_{k+1}^s \, \partial \mathbf{C}_k^T \, \delta t - \mathbf{P}_k \, \partial \mathbf{C}_k^T \, \delta t \\
&\approx \mathbf{P}_{k+1}^s - \mathbf{f}(\mathbf{X}_k, t_k) \, \mathbf{W} \, \mathbf{X}_k^T \, \mathbf{P}_k^{-1} \, \mathbf{P}_{k+1}^s \, \delta t \\
&\quad - \mathbf{L}(t_k) \, \mathbf{Q}(t_k) \, \mathbf{L}^T(t_k) \, \mathbf{P}_k^{-1} \, \mathbf{P}_{k+1}^s \, \delta t - \mathbf{P}_{k+1}^s \, \mathbf{P}_k^{-1} \, \mathbf{X}_k \, \mathbf{W} \, \mathbf{f}^T(\mathbf{X}_k, t_k) \, \delta t \\
&\quad - \mathbf{P}_{k+1}^s \, \mathbf{P}_k^{-1} \, \mathbf{L}(t_k) \, \mathbf{Q}(t_k) \, \mathbf{L}^T(t_k) \, \delta t + \mathbf{L}(t_k) \, \mathbf{Q}(t_k) \, \mathbf{L}^T(t_k) \, \delta t,
\end{aligned}
$$

which gives

$$
\begin{aligned}
\frac{\mathbf{m}_{k+1}^s - \mathbf{m}_k^s}{\delta t} &= \mathbf{f}(\mathbf{X}_k, t_k) \, \mathbf{w}_m \, \delta t \\
&\quad + \mathbf{f}(\mathbf{X}_k, t_k) \, \mathbf{W} \, \mathbf{X}_k^T \, \mathbf{P}_k^{-1} \, \mathbf{m}_{k+1}^s \, \delta t \\
&\quad + \mathbf{L}(t_k) \, \mathbf{Q}(t_k) \, \mathbf{L}^T(t_k) \, \mathbf{P}_k^{-1} \, \mathbf{m}_{k+1}^s \, \delta t \\
&\quad - \mathbf{f}(\mathbf{X}_k, t_k) \, \mathbf{W} \, \mathbf{X}_k^T \, \mathbf{P}_k^{-1} \, \mathbf{m}_k \, \delta t \\
&\quad - \mathbf{L}(t_k) \, \mathbf{Q}(t_k) \, \mathbf{L}^T(t_k) \, \mathbf{P}_k^{-1} \, \mathbf{m}_k \, \delta t \\
\frac{\mathbf{P}_{k+1}^s - \mathbf{P}_k^s}{\delta t} &= \mathbf{f}(\mathbf{X}_k, t_k) \, \mathbf{W} \, \mathbf{X}_k^T \, \mathbf{P}_k^{-1} \, \mathbf{P}_{k+1}^s \\
&\quad + \mathbf{L}(t_k) \, \mathbf{Q}(t_k) \, \mathbf{L}^T(t_k) \, \mathbf{P}_k^{-1} \, \mathbf{P}_{k+1}^s \\
&\quad + \mathbf{P}_{k+1}^s \, \mathbf{P}_k^{-1} \, \mathbf{X}_k \, \mathbf{W} \, \mathbf{f}^T(\mathbf{X}_k, t_k) \\
&\quad + \mathbf{P}_{k+1}^s \, \mathbf{P}_k^{-1} \, \mathbf{L}(t_k) \, \mathbf{Q}(t_k) \, \mathbf{L}^T(t_k) \\
&\quad - \mathbf{L}(t_k) \, \mathbf{Q}(t_k) \, \mathbf{L}^T(t_k),
\end{aligned}
$$

and in the limit $\delta t \to 0$ we get the result in the Algorithm 3.25.

References

Akashi, H. and Kumamoto, H. (1977). Random sampling approach to state estimation in switching environments. *Automatica*, 13:429–434.

Alonso, M. and Finn, E. J. (1980). *Fundamental University Physics, Volume I: Mechanics and Thermodynamics*. Addison-Wesley, 2nd edition.

Alspach, D. L. and Sorenson, H. W. (1972). Nonlinear Bayesian estimation using Gaussian sum approximations. *IEEE Transactions on Automatic Control*, AC-17(4).

Anderson, R. M. and May, R. M. (1991). *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press.

Andrieu, C., de Freitas, N., and Doucet, A. (2002). Rao-Blackwellised particle filtering via data augmentation. In Dietterich, T. G., Becker, S., and Ghahramani, Z., editors, *Advances in Neural Information Processing Systems 14*. MIT Press.

Antonelli, P. L. and Elliott, R. J. (1986). The Zakai forms of the prediction and smoothing equations. *IEEE Transactions on Information Theory*, IT-32(6):816–817.

Applebaum, D. (2004). *Lévy Processes and Stochastic Calculus*. Cambridge University Press.

Åström, K. and Wittenmark, B. (1995). *Adaptive Control*. Addison-Wesley. 2nd edition.

Ballantyne, D. J., Chan, H. Y., and Kouritzin, M. A. (2001). A branching particle-based nonlinear filter for multi-target tracking. In *Proceedings of the Fourth International Conference on Information Fusion*.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2004). *Hierarchical Modelling and Analysis for Spatial Data*. Chapman Hall/CRC.

Bar-Shalom, Y. and Li, X.-R. (1995). *Multitarget-Multisensor Tracking: Principles and Techniques*. YBS.

Bar-Shalom, Y., Li, X.-R., and Kirubarajan, T. (2001). *Estimation with Applications to Tracking and Navigation*. Wiley Interscience.

Barber, D. and Williams, C. K. I. (1997). Gaussian process for Bayesian classification via hybrid Monte Carlo. In Mozer, M. C., Jordan, M. I., and Petsche, T., editors, *Advances in Neural Information Processing Systems 9*. MIT Press.

Beneš, V. E. (1981). Exact finite-dimensional filters for certain diffusions with nonlinear drift. *Stochastics*, 5:65–92.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. John Wiley & Sons.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press.

Blackman, S. and Popoli, R. (1999). *Design and Analysis of Modern Tracking Systems*. Artech House Radar Library.

Bucy, R. S. (1965). Nonlinear filtering theory. *IEEE Transactions on Automatic Control*, 10:198–198.

Bucy, R. S. and Joseph, P. D. (1968). *Filtering for Stochastic Processes with Applications to Guidance*. John Wiley & Sons.

Casella, G. and Robert, C. P. (1996). Rao-Blackwellisation of sampling schemes. *Biometrika*, 83(1):81–94.

Challa, S., Bar-Shalom, Y., and Krishnamurthy, V. (2000). Nonlinear filtering via generalized Edgeworth series and Gauss-Hermite quadrature. *IEEE Transactions on Signal Processing*, 48(6).

Chang, C.-B., Whiting, R. H., and Athans, M. (1977). On the state and parameter estimation for maneuvering reentry vehicles. *IEEE Transactions on Automatic Control*, AC-22:99–105.

Christakos, G. (1992). *Random Field Models in Earth Sciences*. Academic Press.

Cox, H. (1964). On the estimation of state variables and parameters for noisy dynamic systems. *IEEE Transactions on Automatic Control*, 9(1):5–12.

Crisan, D. and Lyons, T. (1997). Nonlinear filtering and measure-valued processes. *Probab. Theory Relat. Fields*, 109:217–244.

Crisan, D. and Lyons, T. (1999). A particle approximation of the solution of the Kushner-Stratonovitch equation. *Probab. Theory Relat. Fields*, 115:549–578.

Daum, F. E. (1984). Exact finite-dimensional nonlinear filters for continuous time processes with discrete time measurements. In *Proceedings of 23rd Conference on Decision and Control*, pages 16–22.

Daum, F. E. (1986). Exact finite-dimensional nonlinear filters. *IEEE Transactions on Automatic Control*, AC-31(7):616–622.

de Freitas, N., Andrieu, C., Højen-Sørensen, Niranjan, M., and Gee, A. (2001). Sequential Monte Carlo methods for neural networks. In *Sequential Monte Carlo Methods in Practice*, pages 359–379. Springer.

Doucet, A., de Freitas, N., and Gordon, N., editors (2001). *Sequential Monte Carlo Methods in Practice*. Springer.

Doucet, A., Godsill, S. J., and Andrieu, C. (2000). On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, 10(3):197–208.

Doucet, A., Vo, B.-N., Andrieu, C., and Davy, M. (2002). Particle filter for multi-target tracking and sensor management. In *Proceedings of the Fifth International Conference on Information Fusion*, pages 474–481.

Einstein, A. (1905). Über die von molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. *Annelen der Physik*, 17:549–560.

Elerian, O., Chib, S., and Shephard, N. (2001). Likelihood inference for discretely observed nonlinear diffusions. *Econometrica*, 69(4):959–993.

Eraker, B. (2001). MCMC analysis of diffusion models with applications to finance. *Journal of Business and Economic Statistics*, 19(2):177–191.

Fearnhead, P. and Clifford, P. (2003). On-line inference for hidden Markov models via particle filters. *J. R. Statist. Soc. B*, 65(4):887–899.

Fong, W., Godsill, S. J., Doucet, A., and West, M. (2002). Monte Carlo smoothing with application to audio signal enhancement. *IEEE Transactions on Signal Processing*, 50(2):438–449.

Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association*, 70(350):320–328.

Gelb, A., editor (1974). *Applied Optimal Estimation*. The MIT Press.

Gelfand, A. E., Dey, D. K., and Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods (with discussion). In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., editors, *Bayesian Statistics 4*, pages 147–167. Oxford University Press.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. R. (1995). *Bayesian Data Analysis*. Chapman & Hall.

Gilks, W., Richardson, S., and Spiegelhalter, D., editors (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall.

Godsill, S. J., Doucet, A., and West, M. (2004). Monte Carlo smoothing for nonlinear time series. *Journal of the American Statistical Association*, 99(465):156–168.

Godsill, S. J. and Rayner, P. J. (1998). *Digital Audio Restoration: A Statistical Model Based Approach*. Springer-Verlag.

Golub, G. H. and van Loan, C. F. (1996). *Matrix Computations*. The Johns Hopkins University Press, third edition.

Gordon, N. (1997). A hybrid bootstrap filter for target tracking in clutter. *IEEE Transactions on Aerospace and Electronic Systems*, 33(1):353–358.

Gordon, N. J., Salmond, D. J., and Smith, A. F. M. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. In *IEEE Proceedings on Radar and Signal Processing*, volume 140, pages 107–113.

Grewal, M. S. and Andrews, A. P. (2001). *Kalman Filtering, Theory and Practice Using MATLAB*. Wiley Interscience.

Grewal, M. S., Weill, L. R., and Andrews, A. P. (2001). *Global Positioning Systems, Inertial Navigation and Integration*. Wiley Interscience.

Guenther, R. B. and Lee, J. W. (1988). *Partial Differential Equations of Mathematical Physics and Integral Equations*. Dover Publications, Inc.

Gunther, J., Beard, R., Wilson, J., Oliphant, T., and Stirling, W. (1997). Fast nonlinear filtering via Galerkin's method. In *American Control Conference. American Automatic Control Council.*

Hayes, M. H. (1996). *Statistical Digital Signal Processing and Modeling*. John Wiley & Sons, Inc.

Hethcote, H. W. (2000). The mathematics of infectious diseases. *SIAM Review*, 42(4):599–653.

Ho, Y. C. and Lee, R. C. K. (1964). A Bayesian approach to problems in stochastic estimation and control. *IEEE Transactions on Automatic Control*, 9:333–339.

Hue, C., Le Cadre, J.-P., and Perez, P. (2001). The (MR)MTPF: particle filters to track multiple targets using multiple receivers. In *4th International Conference on Information Fusion, Montreal, Canada*.

Ionides, E. L. (2004). Inference and filtering for partially observed diffusion processes via sequential Monte Carlo. Technical report, University of Michigan Statistics Department Technical Report #405.

Isard, M. and MacCormick, J. (2001). BraMBLe: A Bayesian multiple-blob tracker. In *Proceedings of International Conference on Computer Vision*, volume 2, pages 34–41.

Jazwinski, A. H. (1966). Filtering for nonlinear dynamical systems. *IEEE Transactions on Automatic Control*, 11(4):765–766.

Jazwinski, A. H. (1970). *Stochastic Processes and Filtering Theory*. Academic Press.

Julier, S. J. and Uhlmann, J. K. (1995). A general method of approximating nonlinear transformations of probability distributions. Technical report, Robotics Research Group, Department of Engineering Science, University of Oxford.

Julier, S. J. and Uhlmann, J. K. (2004a). Corrections to unscented filtering and nonlinear estimation. *Proceedings of the IEEE*, 92(12):1958–1958.

Julier, S. J. and Uhlmann, J. K. (2004b). Unscented filtering and nonlinear estimation. *Proceedings of the IEEE*, 92(3):401–422.

Julier, S. J., Uhlmann, J. K., and Durrant-Whyte, H. F. (1995). A new approach for filtering nonlinear systems. In *Proceedings of the 1995 American Control, Conference, Seattle, Washington*, pages 1628–1632.

Kaipio, J. and Somersalo, E. (2005). *Statistical and Computational Inverse Problems*. Number 160 in Applied mathematical Sciences. Springer.

Kallianpur, G. (1980). *Stochastic Filtering Theory*. Springer-Verlag.

Kalman, R. E. (1960a). Contributions to the theory of optimal control. *Boletin de la Sociedad Matematica Mexicana*, 5(1):102–119.

Kalman, R. E. (1960b). A new approach to linear filtering and prediction problems. *Transactions of the ASME, Journal of Basic Engineering*, 82:34–45.

Kalman, R. E. and Bucy, R. S. (1961). New results in linear filtering and prediction theory. *Transactions of the ASME, Journal of Basic Engineering*, 83:95–108.

Kaplan, E. D. (1996). *Understanding GPS, Principles and Applications.* Artech House, Boston, London.

Karatzas, I. (1988). A tutorial introduction to stochastic analysis and its applications. Lecture notes.

Karatzas, I. and Shreve, S. E. (1991). *Brownian Motion and Stochastic Calculus*. Springer.

Karlsson, R. and Gustafsson, F. (2001). Monte Carlo data association for multiple target tracking. In *IEEE Target tracking: Algorithms and applications, The Netherlands*.

Kermack, W. O. and McKendrick, A. C. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London, Series A*, 115:700–721.

Kitagawa, G. (1996). Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *Journal of Computational and Graphical Statistics*, 5:1–25.

Kloeden, P. E. and Platen, E. (1999). *Numerical Solution to Stochastic Differential Equations*. Springer.

Kouritzin, M. A. and Sun, W. (2005). Rates for branching particle approximations of continuous-discrete filters. *The Annals of Applied Probability*, 15(4):2739–2772.

Kreucher, C., Kastella, K., and Hero III, A. O. (2003). Multitarget tracking using a particle filter representation of the joint multitarget density. In *Proceedings of Signal and Data Processing of Small Targets*, volume 5204.

Krishnamurthy, V. and Elliott, R. (2002). Robust continuous-time smoothers without two-sided stochastic integrals. *IEEE Transactions on Automatic Control*, 47(11).

Kushner, H. J. (1964). On the differential equations satisfied by conditional probability densities of Markov processes, with applications. *J. SIAM Control Ser. A*, 2(1).

Lee, R. C. K. (1964). *Optimal Estimation, Identification and Control*. M.I.T. Press.

Lendasse, A., Oja, E., and Simula, O. (2004). Time series prediction competition: The CATS benchmark. In *Proceedings of IJCNN 2004*.

Leondes, C. T., Peller, J. B., and Stear, E. B. (1970). Nonlinear smoothing theory. *IEEE Transactions on Systems Science and Cybernetics*, 6(1).

Liptser, R. S. and Shiryayev, A. N. (1977). *Statistics of Random Processes I: General Theory*. Springer-Verlag.

Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*. Springer.

Liu, J. S. and Chen, R. (1995). Blind deconvolution via sequential imputations. *Journal of the American Statistical Association*, 90(430):567–576.

MacKay, D. J. C. (1998). Introduction to Gaussian processes. In Bishop, C. M., editor, *Neural Networks and Machine Learning*, pages 133–165. Springer-Verlag.

Mahler, R. P. (2003). Multitarget Bayes filtering via first-order multitarget moments. *IEEE Transactions on Aerospace and Electronic Systems*, 39(4):1152–1178.

Mahler, R. P. (2004). Random sets: Unification and computation for information fusion – a retrospective assessment. In *Proceedings of the Seventh International Conference on Information Fusion*, volume I, pages 1–20.

Maybeck, P. (1979). *Stochastic Models, Estimation and Control, Volume 1*. Academic Press.

Maybeck, P. (1982a). *Stochastic Models, Estimation and Control, Volume 2*. Academic Press.

Maybeck, P. (1982b). *Stochastic Models, Estimation and Control, Volume 3*. Academic Press.

Milton, J. S. and Arnold, J. C. (1995). *Introduction to Probability and Statistics, Principles and Applications for Engineering and the Computing Sciences*. McGraw-Hill, Inc.

Moral, P. D. and Miclo, L. (2000). Branching and interacting particle systems. Approximations of Feynman-Kac formulae with applications to non-linear filtering. In *Sḿinaire de Probabilités XXXIV*. Springer-Verlag.

Murray, J. D. (1993). *Mathematical Biology*, volume 19. Springer.

Neal, R. M. (1996). *Bayesian Learning for Neural Networks*. Springer.

Neal, R. M. (1999). Regression and classification using Gaussian process priors (with discussion). In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., editors, *Bayesian Statistics 6*, pages 475–501. Oxford University Press.

O'Hagan, A. (1978). Curve fitting and optimal design for prediction (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)*, 40(1):1–42.

Øksendal, B. (2003). *Stochastic Differential Equations: An Introduction with Applications*. Springer, 6 edition.

Pikkarainen, H. (2005). *A Mathematical Model for Electrical Impedance Process Tomography*. Doctoral dissertation, Helsinki University of Technology.

Punskaya, E., Doucet, A., and Fitzgerald, W. J. (2002). On the use and misuse of particle filtering in digital communications. In *EUSIPCO*.

Raiffa, H. and Schlaifer, R. (2000). *Applied Statistical Decision Theory*. John Wiley & Sons, Wiley Classics Library.

Rauch, H. E., Tung, F., and Striebel, C. T. (1965). Maximum likelihood estimates of linear dynamic systems. *AIAA Journal*, 3(8):1445–1450.

Ristic, B., Arulampalam, S., and Gordon, N. (2004). *Beyond the Kalman Filter*. Artech House.

Roberts, G. O. and Stramer, O. (2001). On inference for partially observed nonlinear diffusion models using the Metropolis-Hastings algorithm. *Biometrica*, 88(3):603–612.

Sage, A. P. and Melsa, J. L. (1971). *Estimation Theory with Applications to Communications and Control*. McGraw-Hill Book Company.

Särkkä, S., Vehtari, A., and Lampinen, J. (2004a). Rao-Blackwellized Monte Carlo data association for multiple target tracking. In *Proceedings of the Seventh International Conference on Information Fusion*, volume I, pages 583–590.

Särkkä, S., Vehtari, A., and Lampinen, J. (2004b). Time series prediction by Kalman smoother with cross-validated noise density. In *Proceedings of the International Joint Conference on Neural Networks*, pages 1653–1658.

Särkkä, S., Vehtari, A., and Lampinen, J. (2006a). CATS benchmark time series prediction by Kalman smoother with cross-validated noise density. *Neurocomputing*. In press.

Särkkä, S., Vehtari, A., and Lampinen, J. (2006b). Rao-Blackwellized particle filter for multiple target tracking. *Information Fusion*. In press.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association*, 88(422):486–494.

Shiryaev, A. N. (1996). *Probability*. Springer.

Sidenbladh, H. (2003). Multi-target particle filtering for the probability hypothesis density. In *Proceedings of the International Conference on Information Fusion*, pages 1110–1117.

Stengel, R. F. (1994). *Optimal Control and Estimation*. Dover Publications, Inc.

Stone, L. D., Barlow, C. A., and Corwin, T. L. (1999). *Bayesian Multiple Target Tracking*. Artech House, Boston, London.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):111–147.

Storvik, G. (2002). Particle filters in state space models with the presence of unknown static parameters. *IEEE Transactions on Signal Processing*, 50(2):281–289.

Stratonovich, R. L. (1968). *Conditional Markov Processes and Their Application to the Theory of Optimal Control*. American Elsevier Publishing Company, Inc.

Streit, R. L. and Luginbuhl, T. E. (1994). Maximum likelihood method for probabilistic multi-hypothesis tracking. In *Proceedings of Signal and Data Processing of Small Targets*, volume 2235, pages 394–405.

Titterton, D. H. and Weston, J. L. (1997). *Strapdown Inertial Navigation Technology*. Peter Pregrinus Ltd.

van der Merwe, R. (2004). *Sigma-Point Kalman Filters for Probabilistic Inference in Dynamic State-Space Models*. PhD thesis, OGI School of Science & Engineering, Oregon Health & Science University, Portland, OR, USA.

van der Merwe, R., Freitas, N. D., Doucet, A., and Wan, E. (2001). The unscented particle filter. In *Advances in Neural Information Processing Systems 13*.

Van Trees, H. L. (1968). *Detection, Estimation, and Modulation Theory Part I*. John Wiley & Sons, New York.

Van Trees, H. L. (1971). *Detection, Estimation, and Modulation Theory Part II*. John Wiley & Sons, New York.

Vauhkonen, M. (1997). *Electrical impedance tomography and prior information*. PhD thesis, Kuopio University.

Vehtari, A. (2001). *Bayesian Model Assessment and Selection Using Expected Utilities*. Doctoral dissertation, Helsinki University of Technology.

Vehtari, A. and Lampinen, J. (2002). Bayesian model assessment and comparison using cross-validation predictive densities. *Neural Computation*, 14(10):2439–2468.

Vehtari, A. and Lampinen, J. (2003). Expected utility estimation via cross-validation.

In Bernardo, J. M., Dawid, A. P., Berger, J. O., West, M., Heckerman, D., Bayarri, M., and Smith, A. F., editors, *Bayesian Statistics 7*, pages 701–710. Oxford University Press.

Vihola, M. (2005). Random set particle filter for bearings-only multitarget tracking. In Kadar, I., editor, *Proceedings of SPIE*, volume 5809, pages 301–312.

Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, II-13(2).

Vo, B.-N., Singh, S., and Doucet, A. (2003). Sequential Monte Carlo implementation of the PHD filter for multiple-target tracking. In *Proceedings of the International Conference on Information Fusion*, pages 792–799.

Wan, E. A. and van der Merwe, R. (2001). The unscented Kalman filter. In Haykin, S., editor, *Kalman Filtering and Neural Networks*, chapter 7. Wiley.

West, M. and Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models*. Springer-Verlag.

Wiener, N. (1950). *Extrapolation, Interpolation and Smoothing of Stationary Time Series with Engineering Applications*. John Wiley & Sons, Inc., New York.

Wilkie, J. (2004). Numerical methods for stochastic differential equations. *Physical Review E*, 70, 017701.

Williams, C. K. I. and Rasmussen, C. E. (1996). Gaussian processes for regression. In Touretzky, D. S., Mozer, M. C., and Hasselmo, M. E., editors, *Advances in Neural Information Processing Systems 8*. MIT Press.

Zakai, M. (1969). On the optimal filtering of diffusion processes. *Zeit. Wahrsch.*, 11:230–243.

Index